Technology Trends in Kickstarter

Process Book

Matthew Bemtgen, Lauren Kay, Alex Kendall

About our Project

We decided to look at one of the largest crowdfunding platforms, Kickstarter, in order to analyze the various trends in technology innovations over the course of 2016. We decided to focus solely on technology because we feel it is the most rapid growing industry and we are analyzing the past year's data to determine where it is headed in the future.

The Data

We took a lot of time to find a data set that we would be able to make useful analysis from and the data we decided to use is from WebRobots.io. The data is separated into each month, so we will use the files each month in 2016. Within each file, the data is also separated by category, so we will combine all technology Kickstarter projects from each month of 2016 in order to produce the dataset from which we want to analyze. Some of the fields we plan to use from the dataset is the blurb, the city, the size of the goal, the amount raised, the average pledge amount, the number of backers, the country, and whether or not the project is a staff pick. Here is a link to our data: https://webrobots.io/kickstarter-datasets/.

Data Cleaning, Processing & EDA

The data cleaning process was by far the most lengthy and complex part of our project.

1. Used Terminal on Mac to combine the CSV files for each month into one large CSV file. We used a special command to use the header from the first file and then strip the header from subsequent files. At a result of this process, we went from 350 CSV files down to 12.
2. Used R to remove all records that did not have a category that contained Technology. As a result, the number of records in each CSV files decreased by about 80%.
3. Used R to parse out the category field. The categories were saved as links to Kickstarter, and we wanted them to just be simple names. We wrote a script to find and replace the 15 different links with the corresponding category names (e.g. Apps, 3D Printing, etc…)
    a. **Before:** {"urls":{"web":{"discover":"http://www.kickstarter.com/discover/categories/technology/3d%20printing"}},"color":6526716,"parent_id":16,"name":"3D Printing","id":331,"position":1,"slug":"technology/3d printing"}
    b. **After:** 3D Printing

4. Combined all the files back together using Terminal to result in a CSV file that had 200,000 rows and was 480mb in size.
5. We then struggled to figure out how to extract the location from the location column, which was filled with links in a similar way to how the categories were originally defined -- except this time, there were hundreds of locations.
6. We split the CSV files into 10 subset files so that we could open them up in Excel. Since the location was surrounded by quotation marks, we used quotation marks to separate the location into multiple columns and then removed the columns we did not need. Although this was more manual than we would have liked, it allowed us to complete something that had taken us several days to try and figure out in the span of 20 minutes.
    a. **Before:**
    c("{\"country\":\"US\",\"urls\":{\"web\":{\"discover\":\"https://www.kickstarter.com/discover/places/fort-lauderdale-fl\",\"location\":\"https://www.kickstarter.com/locations/fort-lauderdale-fl\"},\"api\":{\"nearby_projects\":\"https://api.kickstarter.com/v1/discover?signature=1460767447.6581750f45ab52141811a166b7abdcb3b866b866&woe_id=2405797\"}},\"name\":\"Fort Lauderdale\",", "Fort Lauderdale, FL\",\"short_name\":\"Fort Lauderdale, FL\",\"id\":2405797,\"state\":\"FL\",\"type\":\"Town\",\"is_root\":false,\"slug\":\"fort-lauderdale-fl\"}")
    b. After: Fort Lauderdale, FL
7. We then used Terminal to combine the subset files back together and complete our master CSV file.
8. In order to do the word bubble from the project blurbs, we used the cSplit function within R to separate the words into separate rows tied to the project ID. Without this, Tableau would have created the

Research Questions

Question 1
Is there a correlation between country and reaching the funding goal? Hypothesis: Percentage of projects fully funded is higher in US.

Question 2
Is the quantity of projects more concentrated in certain geographic areas? Hypothesis: More concentrated on the coasts; San Francisco and New York

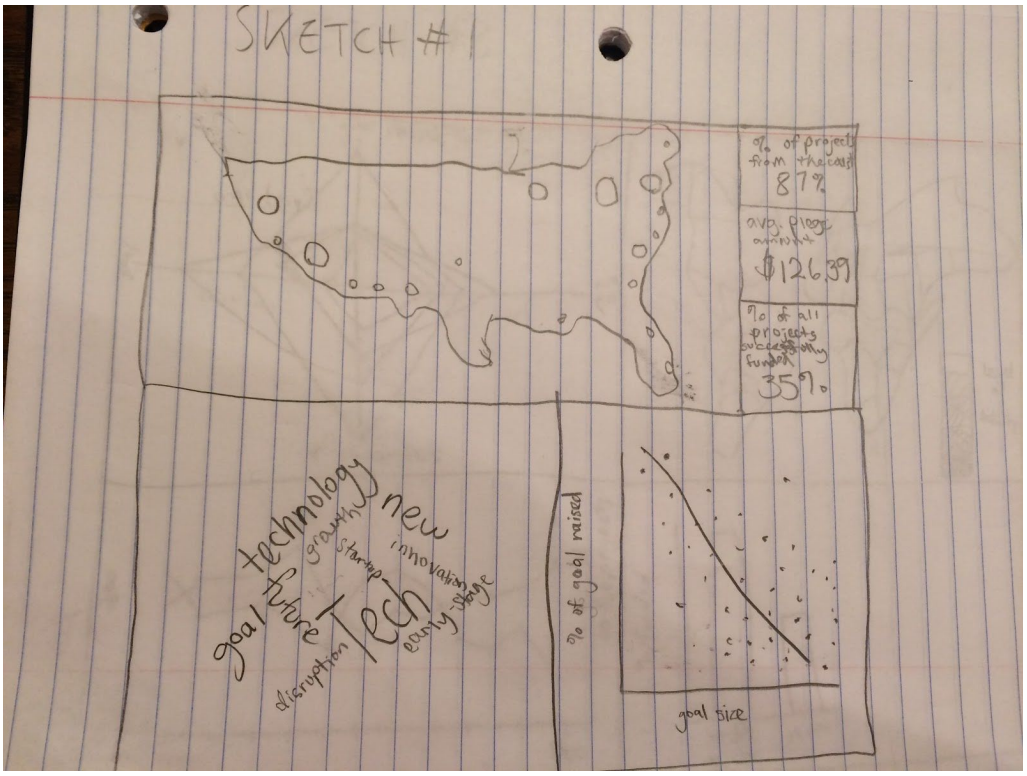
Question 3
What words are most commonly used in the blurb?

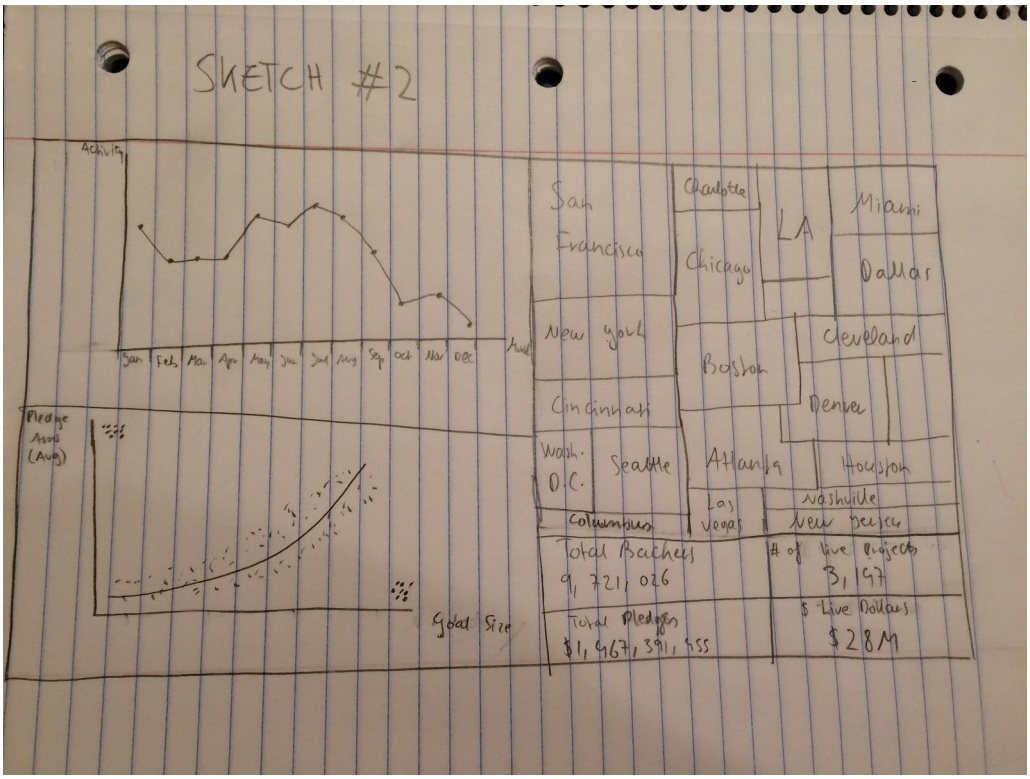
Hypothesis: Goal, Future, Tech, New

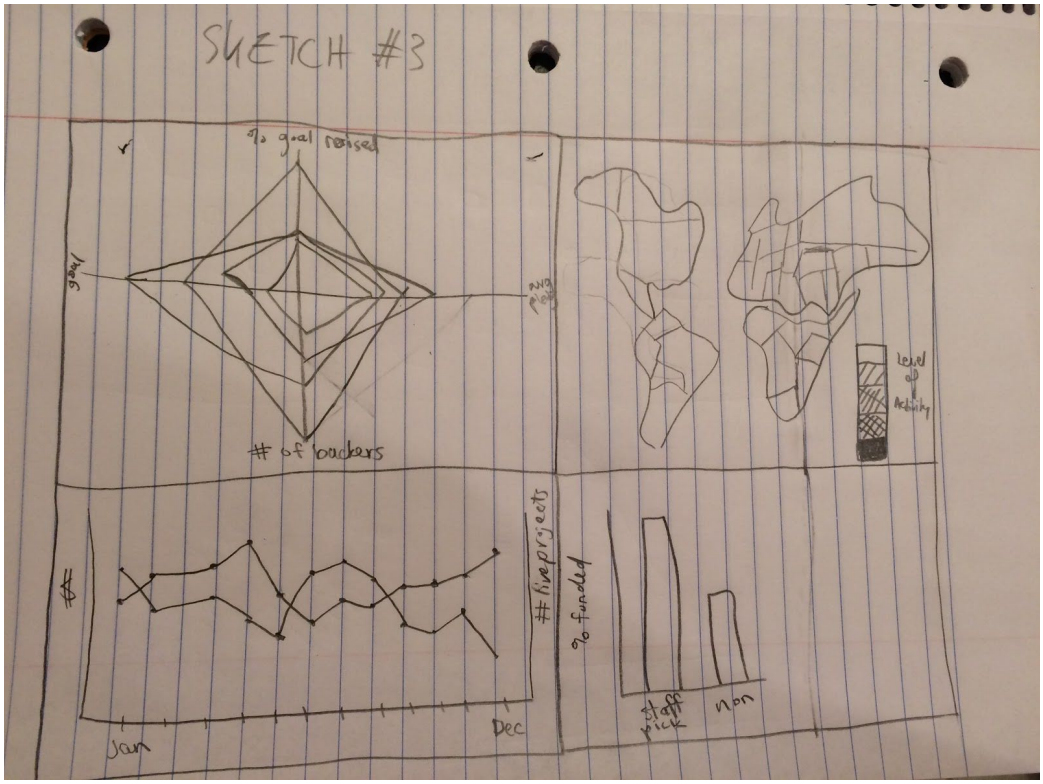
Question 4
Is there a relationship between number of backers and the total pledge amount?
Hypothesis: Higher number of backers tend to have a higher pledge amount.

Question 5
How do staff picks influence the percentage of goal raised?
Hypothesis: Staff picked projects on average raise 90% of their goal.

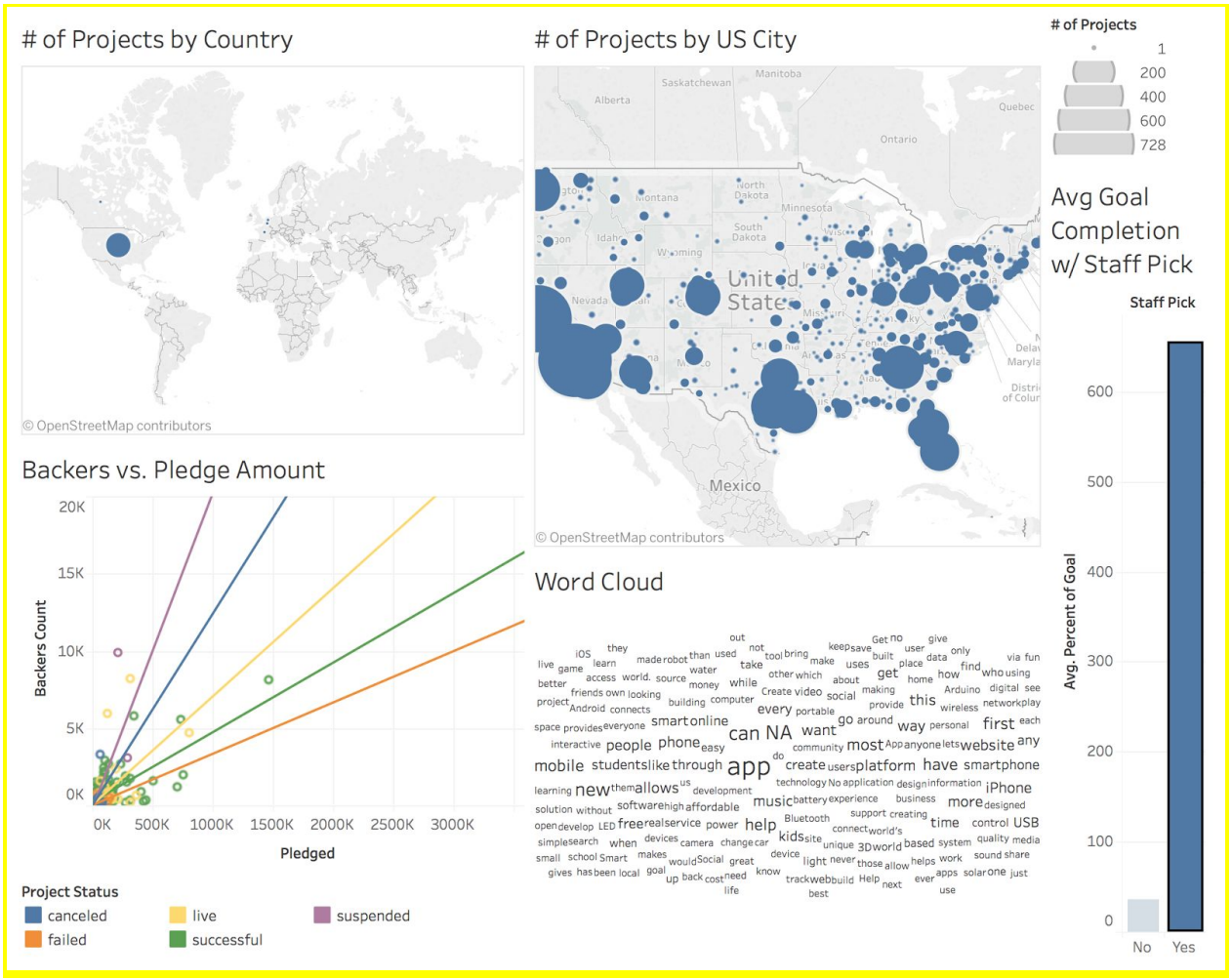
Preliminary Sketches






SKETCH #3
% goal raised
goal
# of backers
Level of Activity
$
Jan
Dec
live projects
% funded
staff pick
non

Final Dashboard



Final Analysis

The first aspect we analyzed is the total number of projects based on the country it originated from. We predicted that the United States would have the most projects, and this proved to be true as shown in this map, where 13,945 projects were started. The country with the next highest number of total projects was Canada with 1,200 projects.

Next we narrowed it down to focus on the city of origin within just the United States. We believed that cities along the coast would have the most projects. Our results correlated with this and some of the cities with the most projects included San Francisco, Los Angeles, San Diego, and Seattle. The largest concentration of projects are cities along the west coast. The west coast of the United States is known for being a hub for startups and innovation.

The next aspect we decided to look at is the blurb which is a brief description of the product or idea. We decided the best way to analyze this is through a word cloud. We predicted that words such as “technology,” “future,” “goal,” and “new” would be larger words in the word cloud. However the words that stand out the most in our word cloud are words such as mobile, app, make, 3D, social, create, design, and online. This shows a trend in which technology is moving towards, shown by the words mobile, app, 3D, social, and online.

We then decided to look at the number of backers against the total amount pledged per project. We predicted that the higher number of backers would result in a higher total pledged amount. As you can see in the graph, the highest correlation between number of backers and total pledge amount are suspended projects that were later restarted, successful projects, and projects that are still live. Suspended projects are set away from the rest, which may mean that the project owner may have decided to take a different approach and suspend the project, later reopening it to get more backers based on the changes they had made. It is not a surprise that the failed and cancelled projects have the lowest correlation between total number of backers and total pledged amount.

Lastly, we wanted to determine what the average percentage of the goal raised for projects that were a staff pick and ones that were not a staff pick. We estimated that on average staff picks would reach 90% of their goal, but our results showed that staff picked projects on average raise 247% of their goal. This is way higher than we had predicted, and a project that is a staff pick is predicted to raise way more that their goal. Projects that are not staff picks raise on average only 12% of their goal.

Even though the large data set gave us a lot of problems, it was very interesting to analyze the data and tell the story through these visualizations using Tableau.