Some Data Tricks and Data Workshop

Log Transformation

A linear regression assumes that the relationship between the dependent variable and independent variable is linear, i.e., we specify the following relationship: $Y = a + bX$.

But sometimes we also take the log-transformation of the linear regression. For example, consider the following relationship:
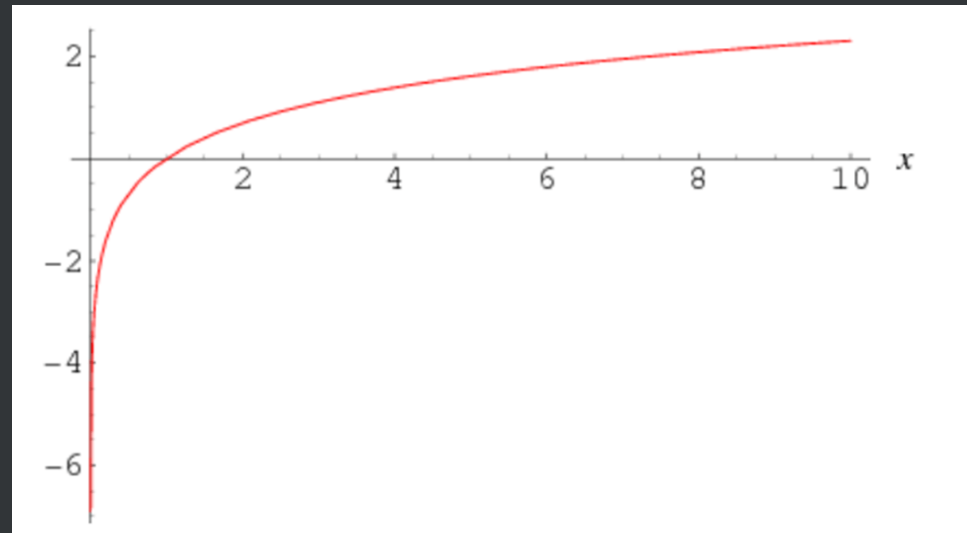
$$\log Y = a + b \times \log X.$$

Here, we typically use the natural logarithms (base is $e \approx 2.718$) in log transformation.

Some basics of the logarithm function:

If $e^a = b$, then $\log(b) = a$, where $e \approx 2.718$.

For instance, we have

- $\log(10) \approx 2.3023$
- $\log(100) \approx 4.6052$

Types of log transformation

- Linear-log models: $Y = a + b \log X$.
- Log-linear models: $\log Y = a + bX$.
- Log-log models: $\log Y = a + b \log X$.

Interpreting your results

- $Y = 0.4 + 11.7 \times \log X$: When X increases by 1%, Y will increase by $11.7 \times 0.01 = 0.17$.
- $\log Y = 1.2 + 0.031 X$: When X increases by 1, Y will increase by 3.1%.
- $\log Y = 2.2 + 0.7 \log X$: When X increases by 1%, Y will increase by 0.7%.

Fixed Effects

Fixed Effects

Consider the following regression: You want to analyze how the sales of a car changes with the color (black, green, yellow, red, white etc.) However, color is not a number, how can you run a regression?

Fixed Effects

You can assign numbers to the color, e.g., `black = 1`, `red = 2`, `white = 3.` However, this is not a great idea: For example, if you find Sales $= 100 + 5 \times$ Color, can you say "when color increases, sales also increase?" This does not make any sense!

Fixed Effects

The solution is to use fixed effects. Instead of creating one single variable for color, we create one variable for each color. For example, for color black, we create the following variable:

$$\text{black} = \begin{cases} 1 & \text{if color is black,} \\ 0 & \text{if color is not black.} \end{cases}$$

Then, we put these variables into our regression equation.

Fixed Effects

```r
data = read.csv("https://ximarketing.github.io/data/fixed_effects.csv")
head(data)
result = lm(sales ~ price + factor(color), data = data)
summary(result)
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       4264.47223    4.83268  882.42   <2e-16 ***
price              -33.88337    0.05926 -571.77   <2e-16 ***
factor(color)red   -77.70687    4.05458  -19.16   <2e-16 ***
factor(color)white 132.51899    3.36567   39.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fixed Effects

The regression output suggests that

$$\text{Sales} = 4264.47 - 33.88 \times \text{Price} + \begin{cases} 0 & \text{if color is black,} \\ -77.71 & \text{if color is red,} \\ 132.52 & \text{if color is white.} \end{cases}$$

Here, black is treated as a benchmark and we compare other colors against this benchmark.

Why do we set a benchmark?

Why color black is set as a benchmark? This is because R adopts alphabetical order, and "black" is before "red" and "white." However, you can change your benchmark as well:

```r
data$color = relevel(factor(data$color), ref = "red")
result = lm(sales ~ price + factor(color), data = data)
summary(result)
```

Changing the benchmark in multinomial logit model:

```r
1 library(nnet)
2 mydata <-
  read.csv("https://ximarketing.github.io/data/multinomial_route_choice.csv")
3 mydata$Choice <- relevel(factor(mydata$Choice), ref = "Freeway")
4 result <- multinom(Choice ~ Flow + Distance +
5                    Seat_belt + Passengers + Age + Male +
6                    Income + Fuel_efficiency, data = mydata)
7 result
```

The Crowdfunding Data

Suppose that you have a great idea, and you believe that your idea can change the world.

But you need resources to implement the idea and turn it into reality. This may cost you hundreds of thousands of dollars.

But you do not have much money yourself. What should you do?

Crowdfunding Platforms

https://www.youtube.com/embed/Vqvomrib6x0?enablejsapi=1

Equity Based Crowdfunding

The backer receives shares of a company, usually in its early stages, in exchange for the money pledged.

Debt Based Crowdfunding

Debt-based crowdfunding is a crowdfunding model used to raise capital by taking loans from several investors (lenders) who expect to be repaid their loan with an added interest over the period that the loan was "used". The entire process takes place through a crowdfunding platform.

Donation Based Crowdfunding

Donation-based crowdfunding is when money is raised to support a good cause. As the name suggests, funding is raised through a crowd of people who decide to donate a certain amount of money to the cause, normally via online platforms specifically designed for the purpose.

Rewards Based Crowdfunding

Rewards-based, or seed, crowdfunding is a type of small-business financing in which entrepreneurs solicit financial donations from individuals in return for a product or service. There are about 19 times as many rewards campaigns as there are for its closely related counterpart, equity-based crowdfunding.

It is closely related to marketing and we focus on it in our class.

" Crowdfunding has the potential to revolutionize the financing of small business, transforming millions of users of social media such as Facebook into overnight venture capitalists, and giving life to valuable business ideas that might otherwise go unfunded.

——The Wall Street Journal

" Besides, isn't this the type of innovation we should be encouraging? Unlike exotic derivatives and super-fast trading algorithms, crowdfunding generates capital for job-creating small businesses

——The New York Times

" *While founders raising cash from a big pool of small amounts of money are benefiting from quick access and the boost of popular interest, they are also forgoing some of the advice and experience of more traditional angel or venture-capital investors*

——The Financial Times

Number of total and repeat Kickstarter project backers from July 2014 to January 2024 (in millions)

Legend: Total backers, Repeat backers
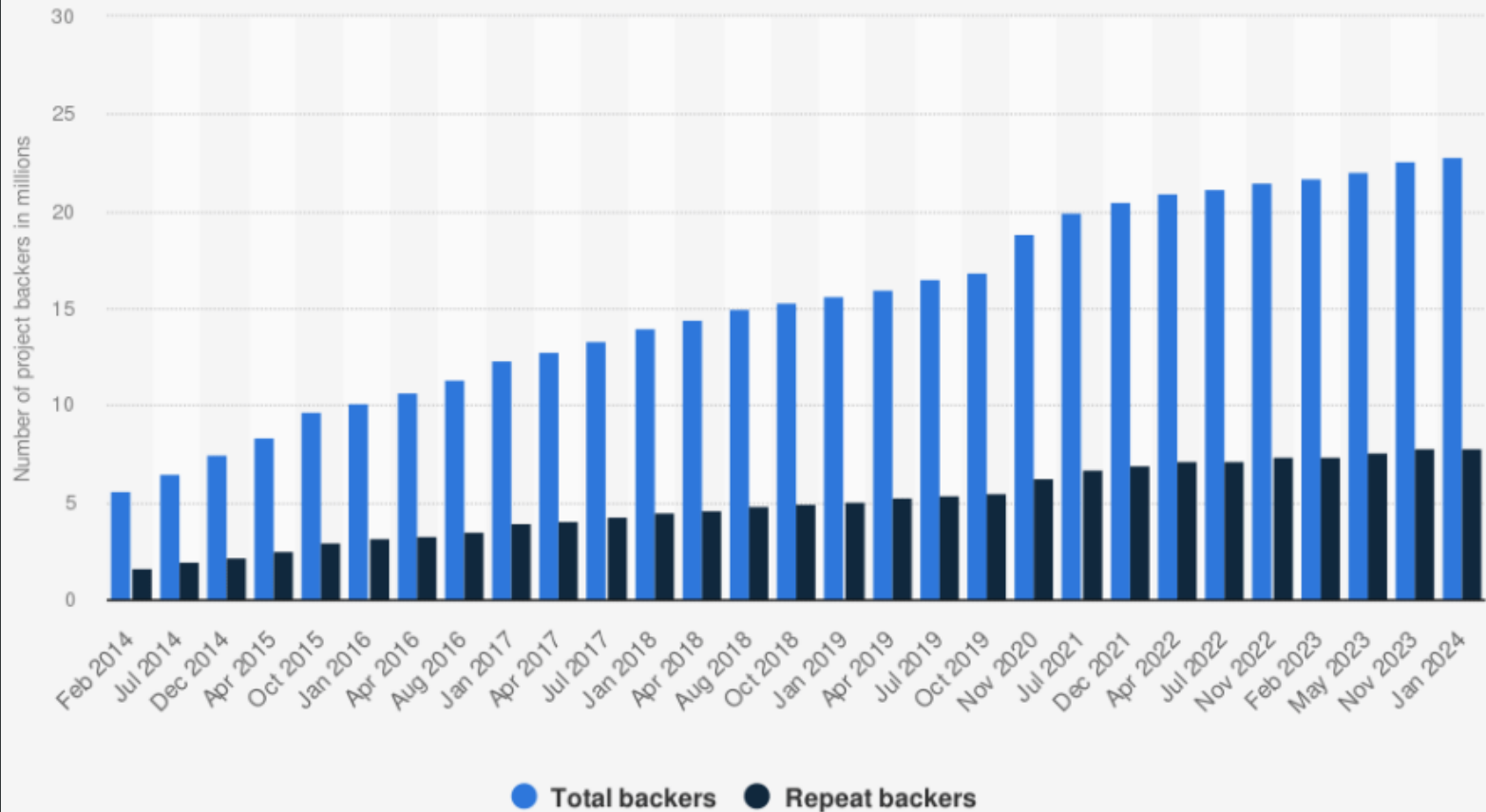
Product Categories in Kickstarter

Kickstarter supports almost all kinds of product categories including Art, Comics, Crafts, Dance, Design, Fashion, Film & Video, Food, Games, Journalism, Music, Photography, Publishing, Technology, and Theater.

Within each category, there are also several subcategories. For example, within the technology category, we have subcategories including gadgets, hardware, DIY electronics, flight, 3D printing, apps, camera equipment, etc.

Pebble Watch

Pebble Watch was a smartwatch developed by the Pebble Technology Corporation. Funding was conducted through a Kickstarter campaign running from April 11, 2012 to May 18, 2012, which raised $10.3 million; it was the most funded project in Kickstarter history, at the time.

Let's visit Pebble Watch's initial crowdfunding webpage to know more about here. Click here to go.

Recall that it is in 2012.

Pebble Watch

Pebble Watch

In 2015, Pebble launched its second generation of smartwatches: the Pebble Time and Time Steel. The devices were similarly funded through Kickstarter, raising $20.3 million from over 75,000 backers and breaking records for the site. See the Kickstarter webpage here.

In 2016, Pebble shut down their subsequent Time 2 series watches and refunded Kickstarter backers, citing financial issues. It was purchased by Fitbit later.

Everyday Backpack

This versatile pack was designed by photographers who felt other camera bags on the market lacked the ability for them to fit all of their other equipment. The bag comes with a number of zippered pockets and waterproof pouches to fit anything you need, as well as versatile handles and anti-theft straps. The campaign's original goal was $500,000, but they ended up with 26,000 backers and $6,565,782 before they packed up and went home.

See the webpage here.

"All-or-Nothing"

Most crowdfunding platforms like Kickstarter strictly implement an "all-or-nothing" policy. That is, the creator (entrepreneur) must set up a target for the project. If the collected fund exceeds the target, the project is successful, and the creator uses the fund to run the project. Otherwise, the project fails, and all the money will be fully refunded to the investors (backers, consumers).

Our Mission

We want to know more about the emerging crowdfunding industry. By doing so, we can

- help entrepreneurs launch better projects and raise more funds from backers;
- help crowdfunding platforms design better features and recommend better projects to backers;
- help the government and public policymakers understand crowdfunding and regulate this industry.

Our Mission

But how to learn about online crowdfunding? You may rely on online materials, conduct surveys, interview experts…

But these methods are really outdated…

Today, you are going to use *real data* to investigate the crowdfunding industry! This is also what data scientists are doing nowadays!

Before Seeing the Data...

Please go to the Kickstarter website, browse a few projects.

If you an entrepreneur trying to launch a successful crowdfunding campaign, what do you want to learn from Kickstarter?

To answer these questions, what data will you collect?

Data

The dataset was scraped from Kickstarter, the largest online crowdfunding website.

- It covers Kickstarter's technology category.
- The data is collected from the following US markets: California, Illinois, Massachusetts, New York, Texas, and Washington State.

Location

The dataset covers projects from six US states: California (CA), New York (NY), Texas (TX), Massachusetts (MA), Washington (WA), and Illinois (IL).
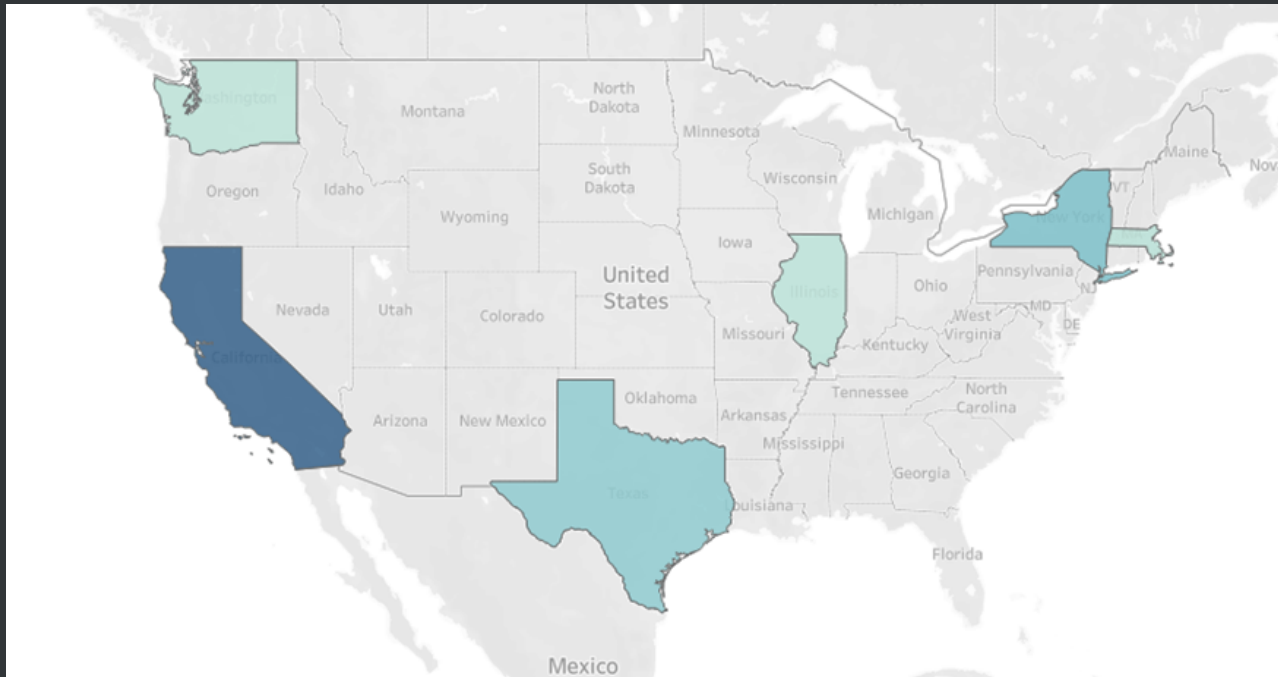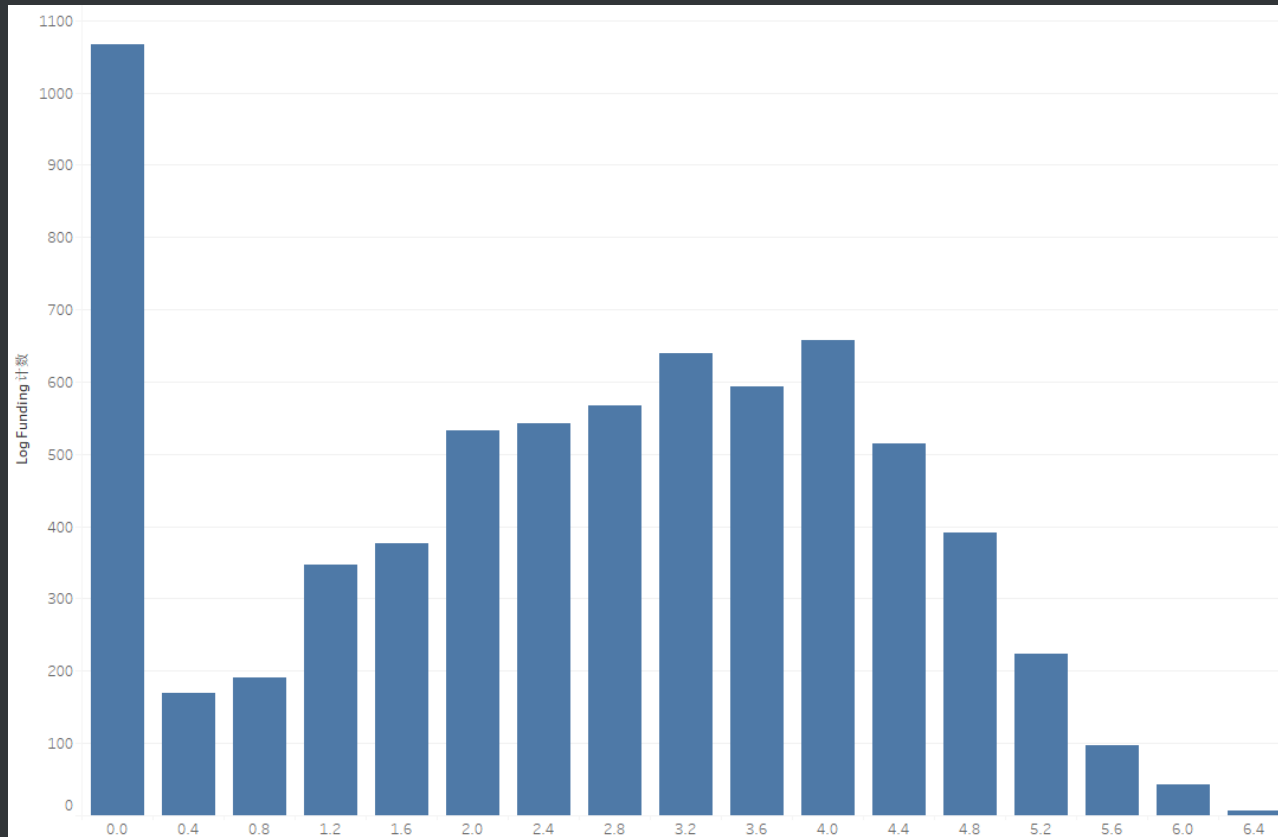
Subtype
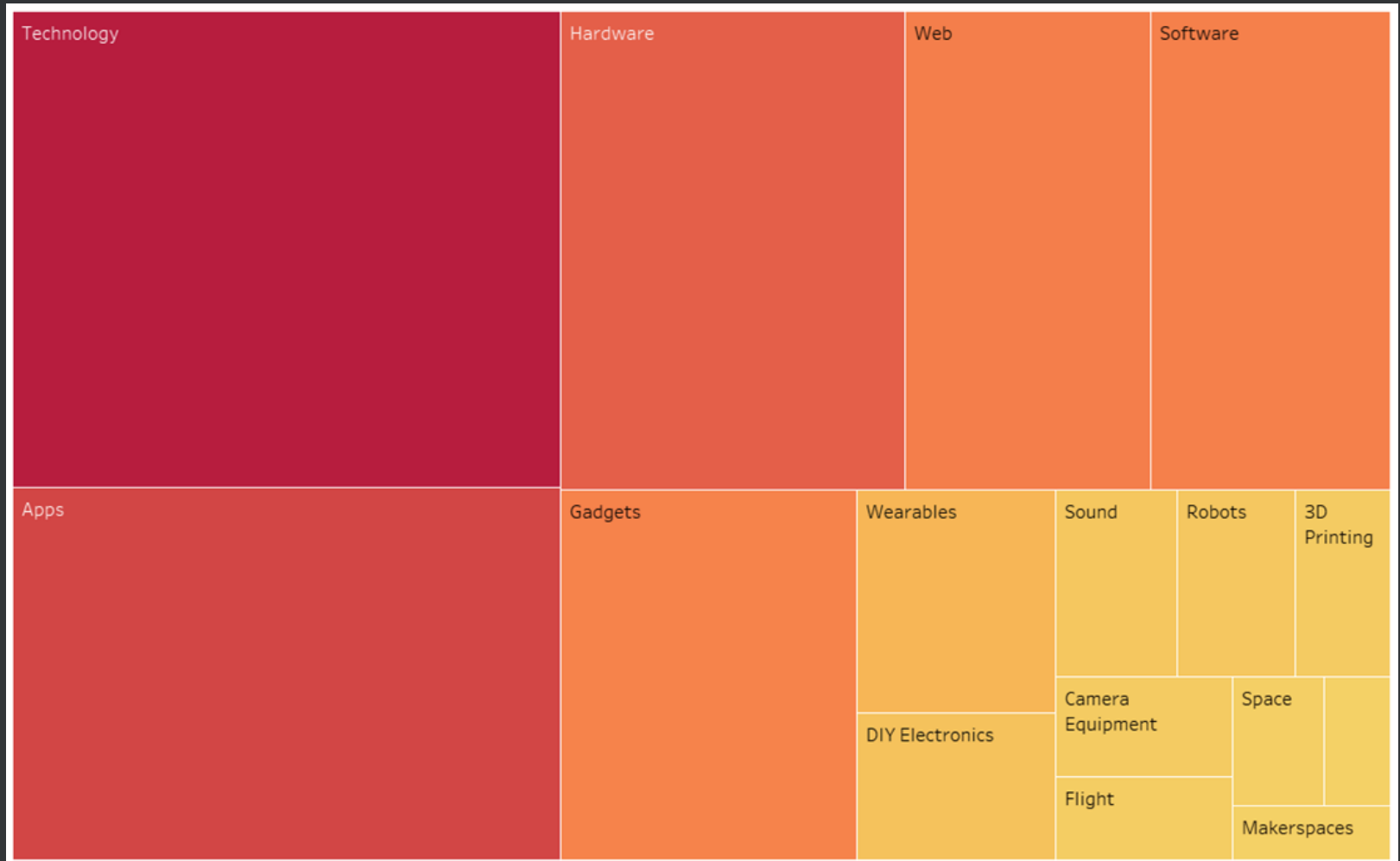
Total Funding Raised
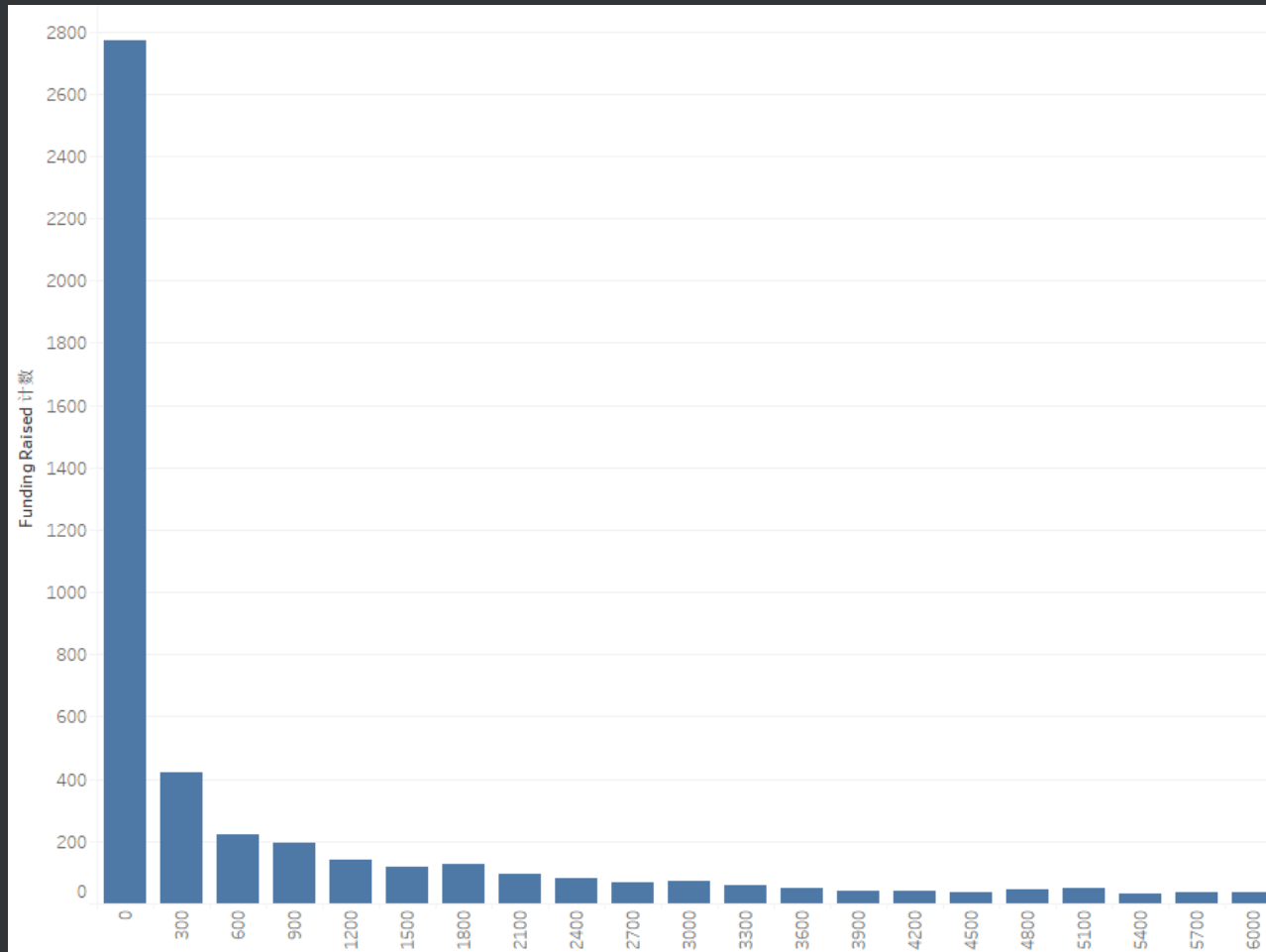
Total Funding Raised

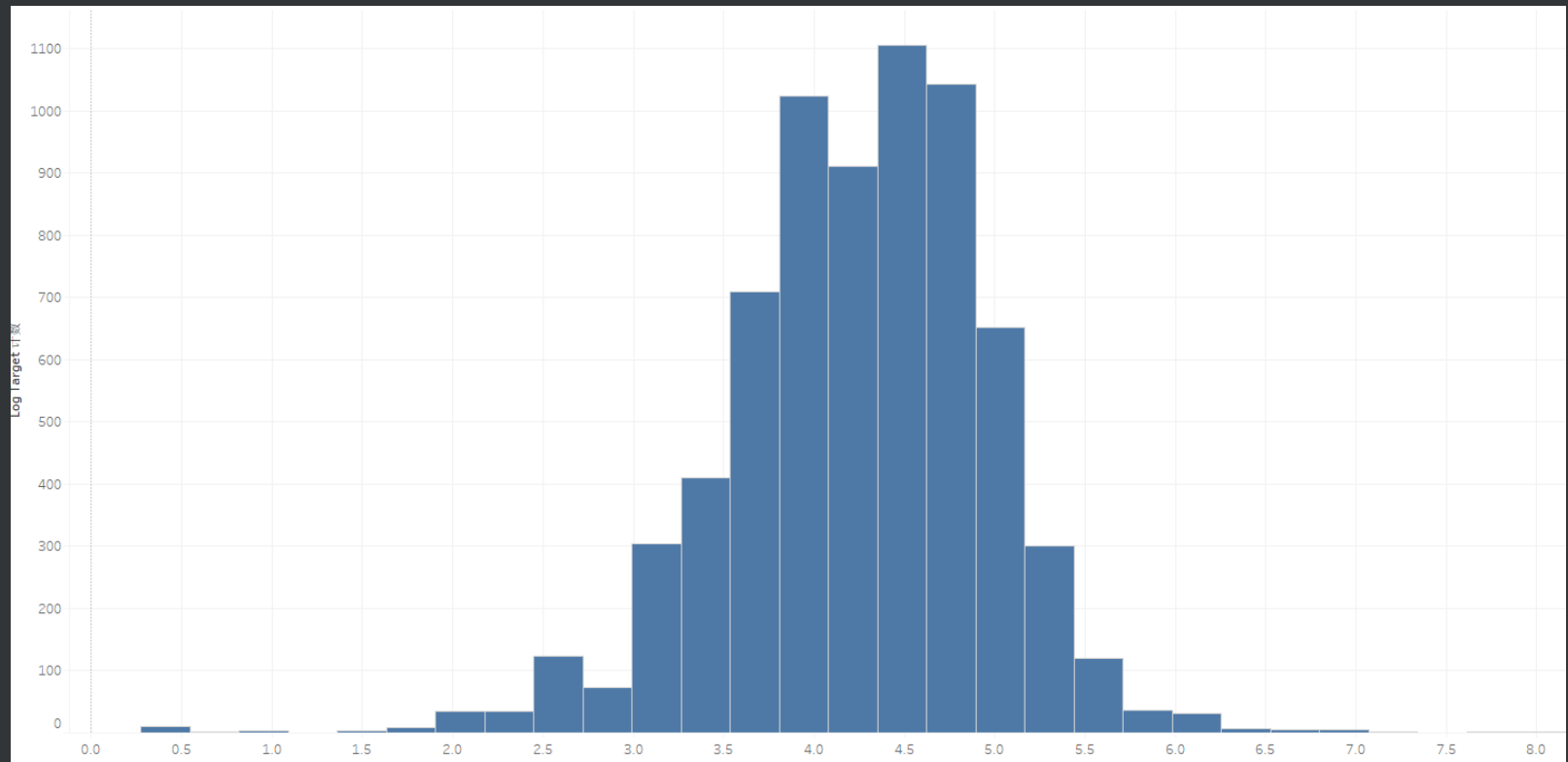The distribution is highly right skewed, and we take log transformation of this variable:

```
1  mydata$LogFundingRaised = log(mydata$FundingRaised + 1)
2  hist(mydata$LogFundingRaised)
```

Why do we use `Funding Raised + 1` here?

# Log Total Funding Raised

Log Target

Other Measures of Project Success

- Outcome: Whether or not the project succeeded. It is a binary variable (1 = success, 0 = failure).
- Backers: Number of people supporting the project. If you divide funding raised by the number of backers, you will get the average fund contributed by a backer.

Entrepreneurs' Personal Background

- Created: Number of projects created by the same entrepreneur in the past. For example, 4 means the same entrepreneur had already created another 4 projects on Kickstarter.
- Backed: Number of projects backed by the same entrepreneur in the past (i.e., the entrepreneur supporting others' projects on Kickstarter).
- FbNumber: Number of Facebook friends the entrepreneur has.

Duration (Days)
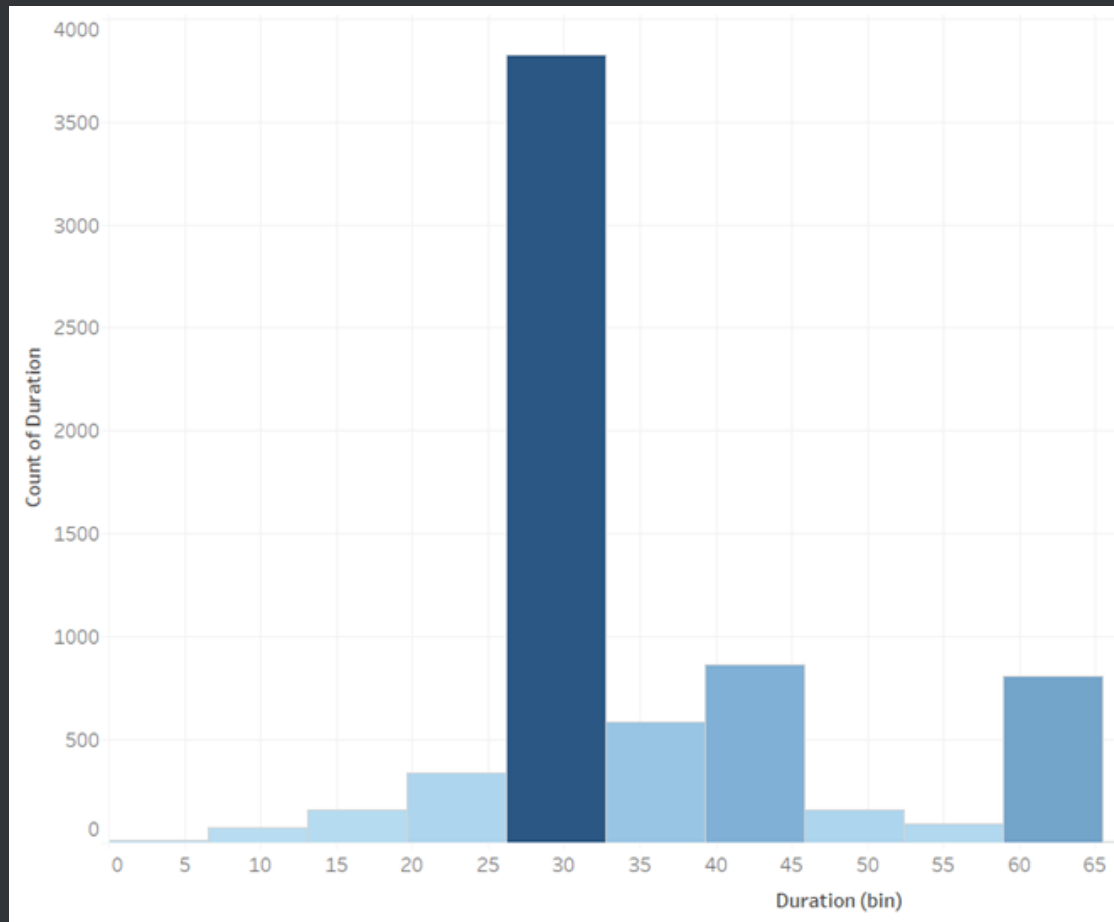
Gender

In the dataset, we have three genders: males (M), female (F), and unknown (U). The gender is obtained by analyzing the creators' first name. Unknow refers to the case in which the name cannot be identified (e.g., a team name such as "marketing").

Video Related Variables

- Video: Whether or not the project has a video. In the dataset, 76% of the projects have a video. Here, 1 means has video and 0 means no video.
- Human: Whether or not the project's video features human-beings (usually the entrepreneurs themselves). 1 means has human and 0 means non-human. This variable is set to 0 is the project does not have a video.
- Computer: Whether or not the project's video features a computer. 1 means has computers and 0 means no computers. This variable is set to 0 is the project does not have a video.
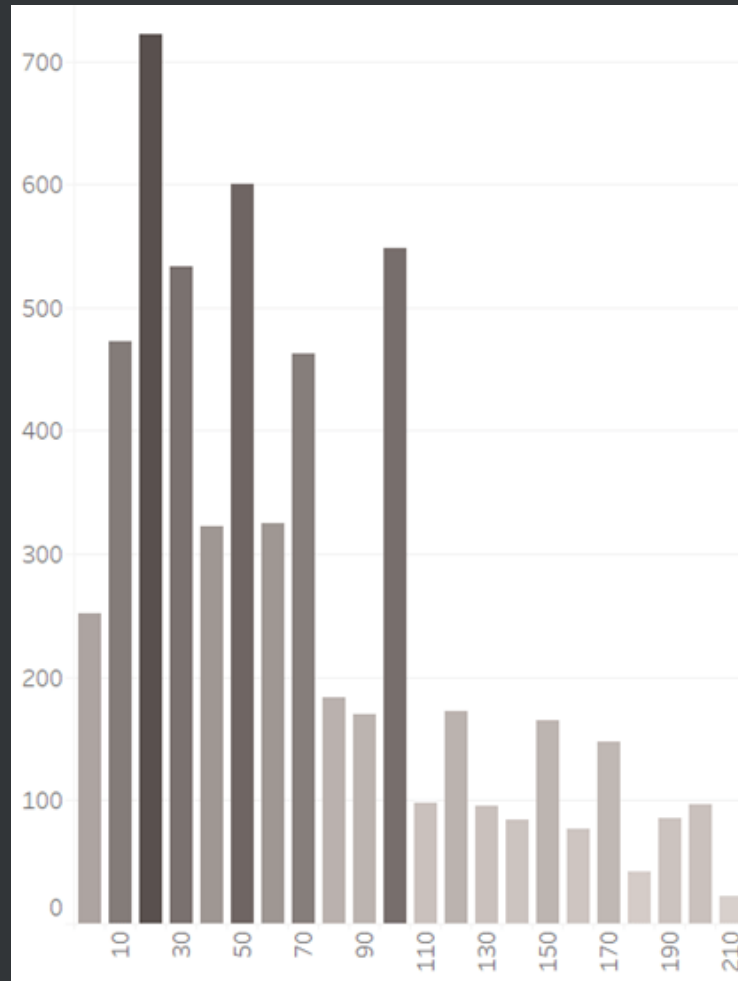
Video Length (Seconds)
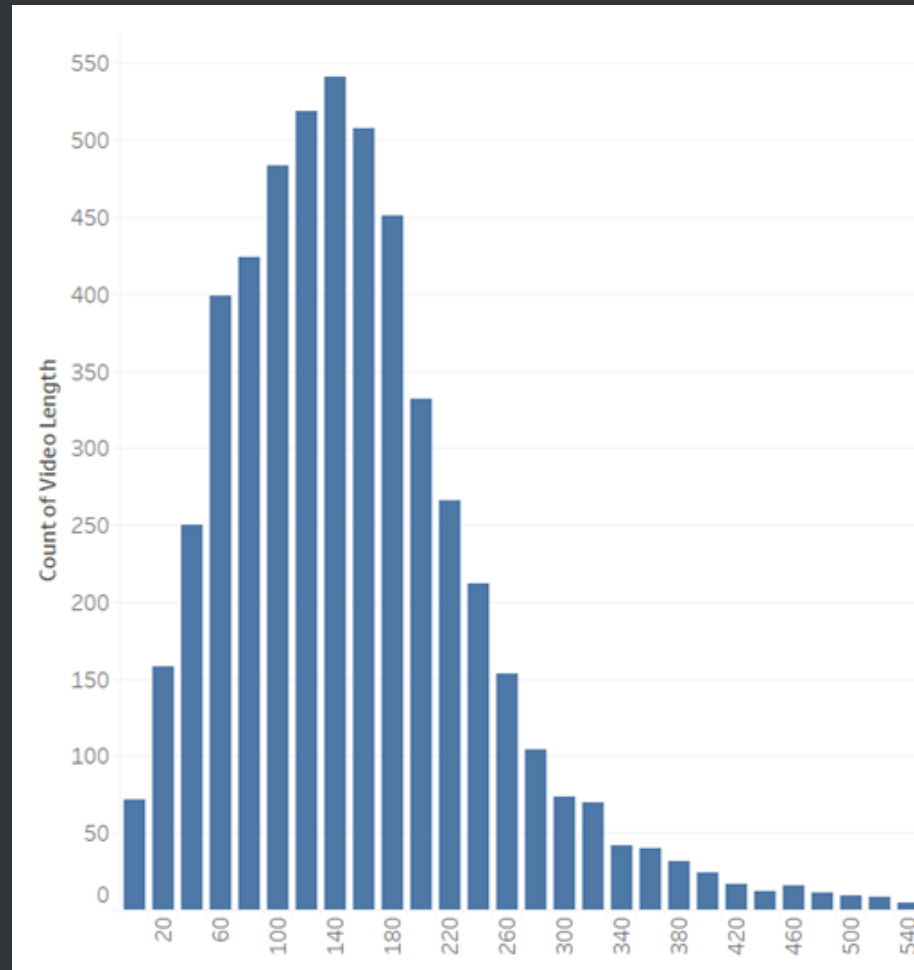
The value is 0 if there is no video at all.
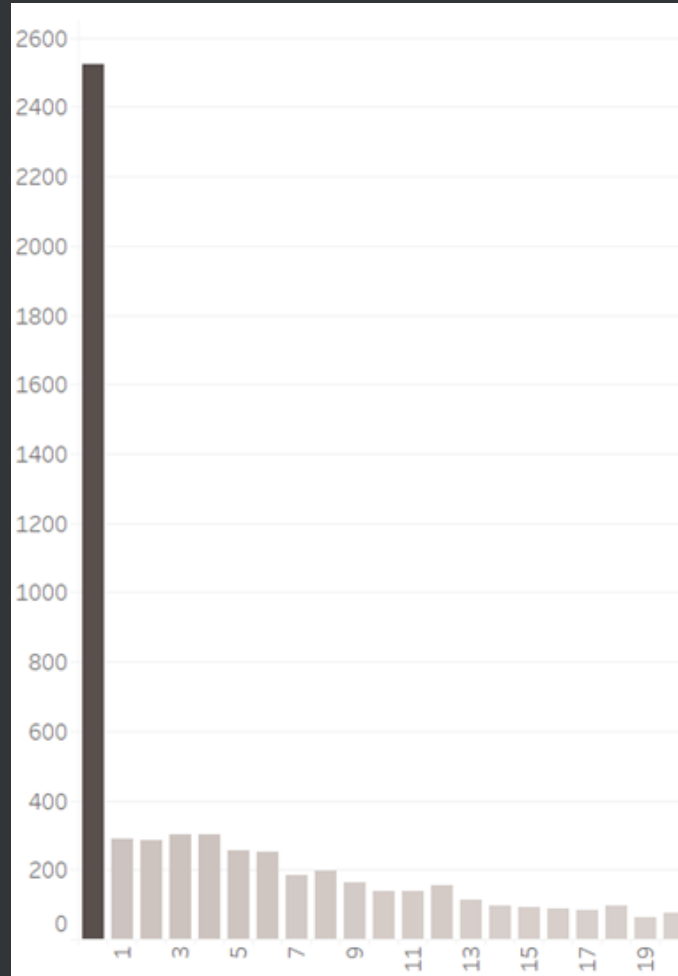
Number of Products
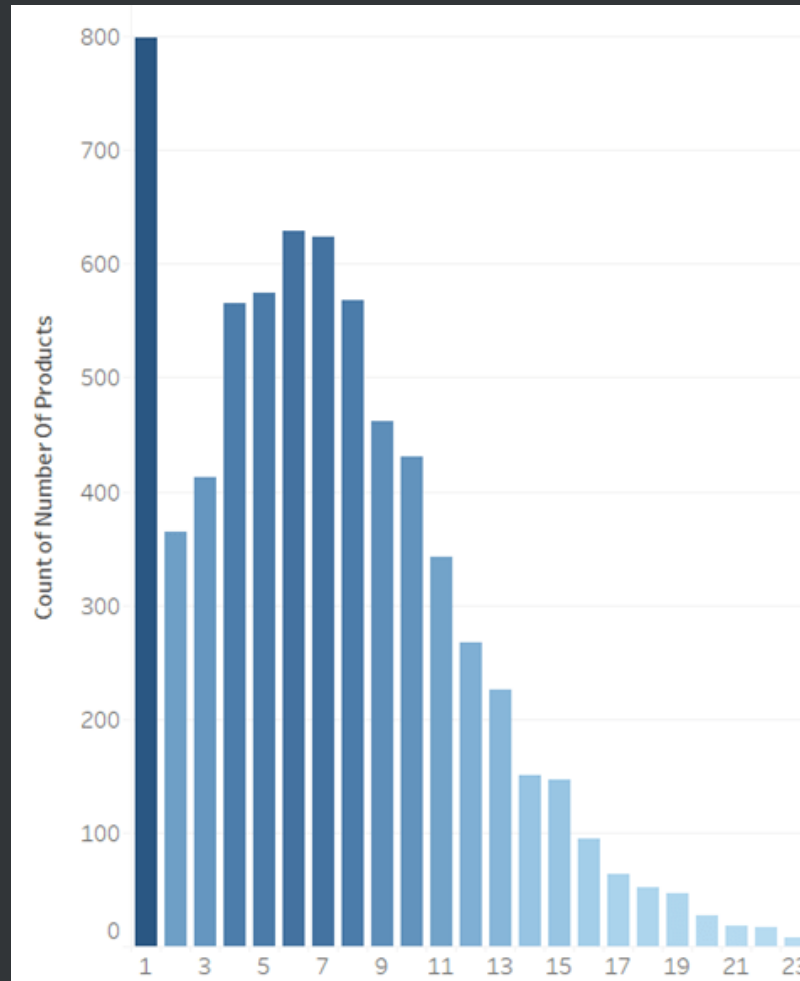
How many products are offered in the project.

Price

Median price among all product offerings.

Photo Number

The number of photos in the campaign page.

Comments

Number of comments posted by consumers.

Audio Variables

- Energy: Whether or not the audio pitch sounds energetic. A large number of an energetic audio pitch.
- Content: Whether or not the audio pitch shows signs of content.
- Upset: Whether or not the audio pitch shows signs of upset.
- Angry: Whether or not the audio pitch shows signs of anger.
- MaxAmpVol: The max sound volume. A greater number means louder sound.

What should I do?

Use the data to provide recommendations for the platform or the entrepreneurs. You can focus on anything that can be helpful for the platform or the entrepreneurs:

1. How do male and female entrepreneurs behave differently on Kickstarter? (e.g., compared to females, males may be too aggressive in setting high targets.)
2. Which type of video is most productive in terms of generating funds? (e.g., is having a lengthy video always beneficial?)
3. What makes a successful crowdfunding project?

What should I do?

Each group should only ask one big or two small research questions in your project. Quality beats quantity. Choose the right data analysis methods and come up with a good answer to your questions, with implications for platforms or entrepreneurs.

What should I do?

A big question needs to be answered with multiple analyses. For example, you may need to run a few regressions to answer a big question.

A small question can be answered within one or two steps.

Sample Analysis

```r
1  mydata <-
   read.csv("https://ximarketing.github.io/class/Kickstarter-
   Project.csv", fileEncoding = "UTF-8-BOM")
2  summary(mydata)
3  mydata$LogTarget = log(mydata$Target + 1)
4  mydata$LogFundingRaised = log(mydata$FundingRaised + 1)
5  result <- lm(LogFundingRaised ~ LogTarget + factor(Gender),
   data = mydata)
6  summary(result)
```

Deliverables

To save your time and my time, you only need to submit a few pages of slides (no more than 12 slides main text + no more than 6 slides appendix) to Moodle covering your research question(s), data analysis (e.g., regression equations), findings, and implications.

Deadline: Dec 19, 2024 (one week from today)

Class A: 12:30 pm, Class B: 5:00 pm, Class C: 9:30 pm

No reports. No presentations. Nothing else.