**Logo

Description automatically generated with low confidence**

**ALY 6010: Probability Theory and Introductory Statistics**

**Exploratory Data Analysis of Kickstarter**

**Submitted by: Sakshi Mehta**

**Submitted To: Prof. Amin Karimpour**

**Introduction**

This dataset is all about the Kickstarter campaign. Kickstarter is for start-up companies that are looking to launch their first products. Websites like Kickstarter and Indiegogo provide a platform for millions of creators to present their innovative ideas to the public.

So here people publish their product and have a goal $ to achieve so we can analyze which companies reach the goal and what factors help them reach the goal. This seems quite interesting with all the information about the projects that have succeeded, failed, canceled, or are live.

After applying exploratory data analysis on this dataset, interesting insights and relations can be drawn between backers, funding received, and the goal set.

The dataset used here is from Kaggle with the name – Kickstarter.

The dataset has multiple variables (375765 rows and 15 columns).

It has 15 variables:

-ID (numeric)

-Name (Project name - string)

-Category (Subcategory - string)

-main\_category (string)

-currency (currency used - string)

-deadline (date)

-goal (numeric)

-launched (Date)

-pledged (numeric)

-state (string)

-backers (numeric)

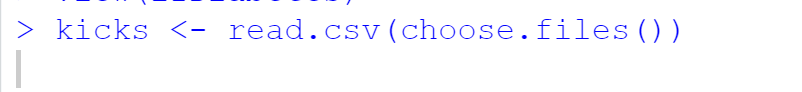
-country (string)

-usd pledged (numeric)

-usd\_pledged\_real (numeric)

-usd\_goal\_real (numeric)

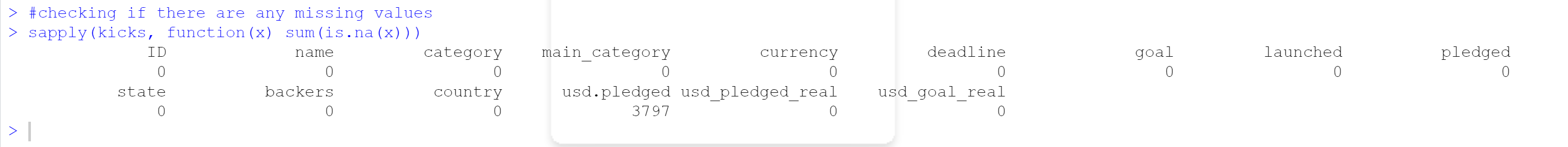
**Loading the dataset and naming the dataframe as – “kicks”**

****

**Table

Description automatically generated**

**Checking if there are any missing values in any of the columns**

****

We can see that the column – ‘usd.pledged ’ has maximum missing values, so that column should be dropped and the column – “usd\_pledged\_real” is renamed with “usd. pledged”.

**Data Analysis**

**Barplot that examines the popularity of projects**

Chart, bar chart, histogram

Description automatically generated

Through this bar plot, it is clear that Film and Video appear to be the most popular category and dance is the least popular.

**Number of projects based on subcategory**

Chart, bar chart, histogram

Description automatically generated

Analyzing the same data for subcategories. It is observed that Product Design is the most popular coming from the category of Design.

**Top projects that got maximum funding.**

Text

Description automatically generated

From the perspective of the backers, it is the most funded and the most popular. The data above shows the top 15 highest funded projects.

**Projects that got maximum backing**

Text, letter

Description automatically generated

Similarly, above is the list of the top 15 most backed projects.

**Total funding grouped by categories.**

Chart, bar chart, histogram

Description automatically generated

After aggregating the amount of funds pledged for each category, we get the total amount pledged for each category. We see that Games, Design, and Technology are the highest-grossing categories so far.

**Distributions of amount pledged for individual projects.**

Chart, bar chart

Description automatically generated

Here, we observe that there are projects that received little to no funding and therefore the box plot looks squished near the bottom. Moreover, it is observed that design, dance, and theater have a high median amount pledged as compared to other categories. Technology has a low median pledged, despite having a high amount pledged, suggesting that there are a lot of outliers.

**Success and failure of projects**

Chart, bar chart

Description automatically generated

We can see that more projects have been failed than succeeded.

**Number of projects based on years**

Chart, bar chart

Description automatically generated

The graph above explores the fact that 2015 was the year with the highest number of projects. It seems the number of projects per year has been gradually decreasing from then, with the number of projects in 2016 and 2017 being lower than both that of 2015 and 2014.

**Does project length affect success rate?**

Kickstart has a rule that the maximum project duration should be 60 days and it is recommended that projects should be set to 30 days or less as if it has not been funded within 30 days then they are not likely to be funded by their deadline either.

Chart, scatter chart

Description automatically generated

It seems that overall, projects exceeding 30 days have a lower success rate than projects lasting fewer than 30 days, so there is some truth to what Kickstarter says.

Diagram

Description automatically generated

Projects seem to primarily originate from North America, Europe, and Oceania, with Japan standing out from the rest of Asia. The United States, and to a lesser extent the United Kingdom, have the highest number of projects**.**

**Summary**

After performing an Exploratory Data Analysis (EDA) on the dataset about Kickstarter several patterns can be seen and used for predictive analysis. Using various visualization techniques and identifying which one can be best used for gaining insights into different relations between variables, useful decisions can be made.