Africa, Zachary Aldridge L.  
Fermin, Lanz Railey A.

**INITIAL DATASET REPORT**

**Overview of the Dataset**

| **Name** | For this project, we combined two datasets.  Kickstarter Projects  2016 Country Economic Profiles |
| --- | --- |
| **Sources** | *Kickstarter Projects*:It is an open-access dataset available via Kaggle, uploaded by Mickaël Mouillé.  *2016 Country Economic Profiles*: The organized dataset of basic economic indicators like population, poverty rate, human development index (HDI), life expectancy (in years), expected and mean schooling years, and the gross national income (GNI) were lifted from a separate Kaggle dataset linked with the analysis of crowdfunding successes for another platform, named Kiva. No information about the creator of this dataset is provided, aside from the fact that this was uploaded by a user named Beluga.  Both datasets can be used under the license CC BY-NC-SA 4.0. |
| **Description** | This dataset provides valuable insights for crowdfunding platform administrators, project creators, and potential backers by offering a data-driven approach to evaluate project viability and optimize the crowdfunding process.  Integrated with economic profiles of the countries where each project originated, we aim to see how these factors may affect a project’s success or failure within the Kickstarter Platform. |

**Methodology**

| **Collection Process** | According to the uploader of this dataset, the data were collected from the [Kickstarter Platform](https://www.kickstarter.com/help/stats) website.  Following this link, it is stated that the website is automatically updated at least once a day with the raw data behind Kickstarter. Several metrics, which are used in the dataset, like funding success status, pledged amount (in USD), and the projects relative to their launch and deadline dates are recorded.  On the other hand, the economic indicators were extracted from the following sources:   * Population – from the World Population Dashboard of the UN Population Fund. * Poverty rates – from the World Population Review of the Organization for Economic Cooperation and Development. * HDI – from the Human Development Reports as stated by the UN Development Programme.   Note: All of the associated metrics with HDI that were used in this project, like life expectancy, expected and mean years of schooling, and the GNI were extracted from the same source. |
| --- | --- |
| **Tools Used** | MS Excel was used to do some preliminary observations on the dataset, and to ensure that the dataset can be read as a CSV.  Jupyter Notebook was used for the bulk of the exploration, data preprocessing, and other tasks done in this phase. |
| **Data Size** | The kickstart\_econ.csv is 64.3 MB.  It contains 323 750 rows and 24 columns. |

**Data Quality Assessment**

*Data Accuracy*

In the current context of our dataset which entails crowdfunding project information and country economic profiling, data accuracy must be ensured so that the assessment we will do in the latter parts of the project is reliable. Given that the sources for the project were presented, which are listed above in the *Collection Process*, it gives a minimum guarantee that the information embedded within the dataset is factual and realistic.

*Data Completeness*

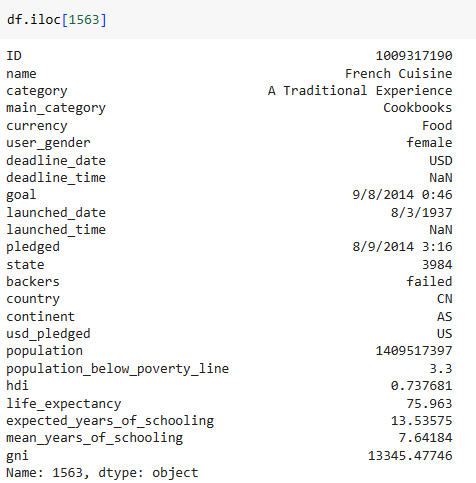
Upon checking Kaggle's criteria on ‘Completeness’, we saw that both datasets (for Kickstarter Platform and 2016 Country Economic Profiles) only reported 33% ratings, garnering disapproval on the basis of Source/Provenance and Update Frequency.

Despite this, we do not see it as a major problem in using the data since the datasets have stated their sources, which compensates for the lack of updates since they were uploaded on Kaggle. Nevertheless, in the data preprocessing, we will perform the necessary tasks to ensure that the dataset is complete (e.g., checking for NaN or NA values and dropping them, and attempting to interpolate missing values, if there are any). A quick observation of the CSV file using MS Excel revealed some rows that needed to be removed due to mismatches with columns and/or blank values.

*Data Consistency*

Data consistency is important to maintain so that, again, the reliability of our analysis after modeling would remain intact. Although no major inconsistencies were detected in the whole dataset, there are occasional exceptions wherein a row would have values that are not matched with the column it is referring to.

For example, it can be seen in Figure 1 below that Row 1563 has a value of ‘Food’ in currency, which should not be the case. It also has a ‘USD’ value in deadline\_date and a NaN in deadline\_time. There are no quick fixes for this, and given the sporadic nature of such errors, we figured that the wise decision would be to remove them altogether.



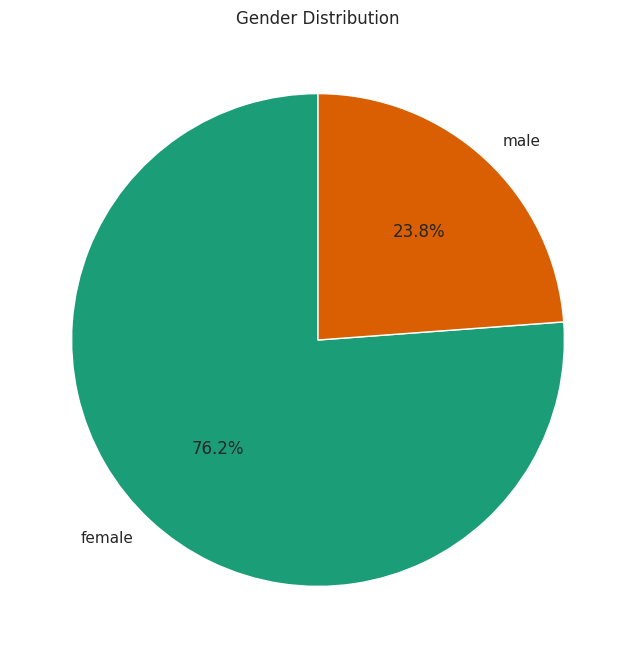
**Figure 1.** An example of a row that has mismatches on the expected value format and on the actual entry.

*Labels and Label Distribution*

In this section, we describe the labels that are present in our dataset.

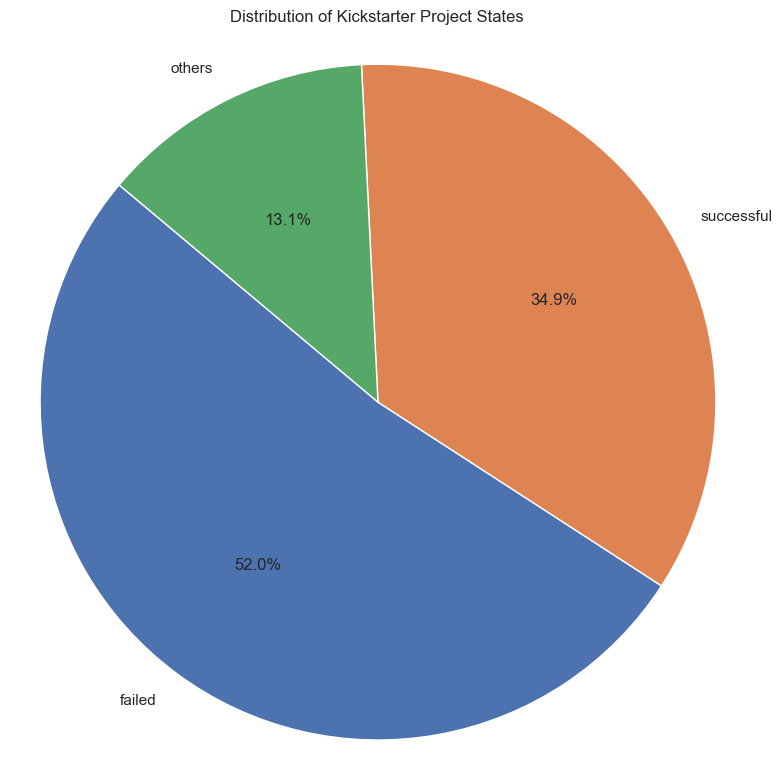
| **Label** | **DataType** | **Description** |
| --- | --- | --- |
| ID | integer | Unique ID for each crowdfunding project. |
| name | string | Name of project. |
| category | string | Fine classification of the project nature. |
| main\_category | string | General classification of the project nature. |
| currency | string | Currency of the crowdfunding project. |
| user\_gender | string | Binary data. Either ‘male’ or ‘female’. |
| deadline\_date | date | Deadline date. |
| deadline\_time | time | Deadline time at the specified date. |
| goal | integer | Target amount to be raised by the project. |
| launched\_date | date | Launch date of project. |
| launched\_time | time | Launch time at the specified date. |
| pledged | float | How much the project currently has raised. |
| state | string | Current state of project (can be success, failed, canceled, suspended, and live; but dropped the latter three to focus on success/fail rates). |
| backers | integer | Number of supporters. |
| country | string | Country of origin. |
| continent | string | Continent of country. |
| usd\_pledged | float | Total amount pledged in USD. |
| population | integer | Population of the country. |
| population\_below\_poverty\_line | float | Poverty incidence rate of the country. |
| hdi | float | HDI rating of the country. |
| life\_expectancy | float | HDI metric. |
| expected\_years\_of\_schooling | float | HDI metric. |
| mean\_years\_of\_schooling | float | HDI metric. |
| gni | float | HDI metric. |

We display some notable and interesting insights with regard to the columns of our dataset.



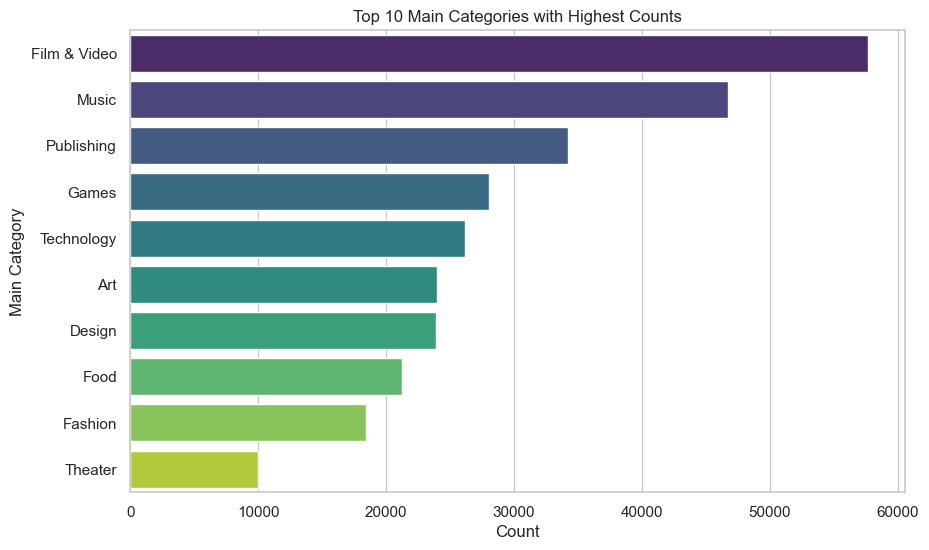
**Figure 2.** Gender distribution of the users who posted crowdfunding projects.

As we can see, approximately three out of four crowdfunding projects in Kickstarter Platform are initiated by a female user.



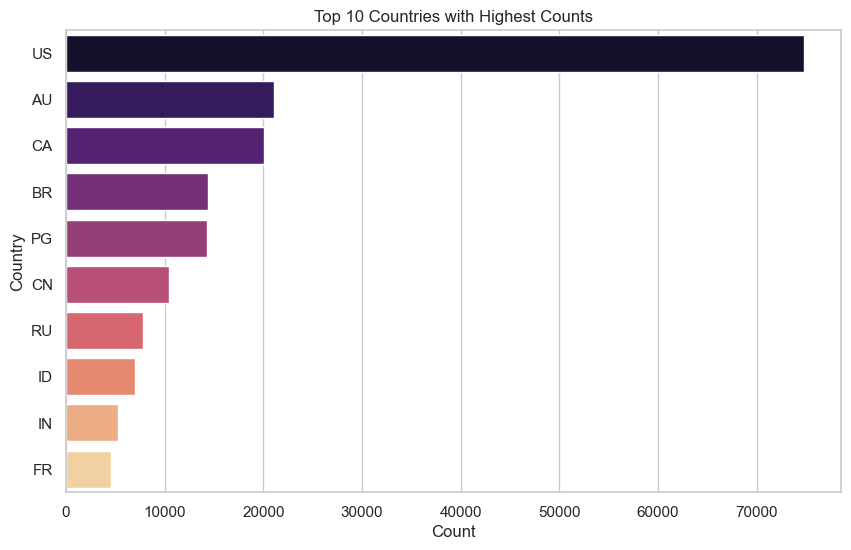
**Figure 3.** Distribution of Kickstarter Platform project states.

From the graph, it can be seen that the ‘Failed’ status of crowdfunding projects constitute the majority. The 34.9% success rate of projects in our dataset remains true to the real-life statistic that around 40% of crowdfunding projects across all possible platforms become successful. Meanwhile, the remaining 13.1% of miscellaneous labels (canceled, suspended, live, etc.) would be dropped in data preprocessing.



**Figure 4.** Top 10 entries in main\_categories with the highest number of frequency.

Film & Video is the most common nature of the crowdfunding projects in Kickstarter, with more than 50 000 entries under its name. Music, Publishing, Games, Technology, Art, Design, Food, Fashion, and Theater complete the labels with the most frequencies. Using df.nunique() we note that there are 120 existing main categories.



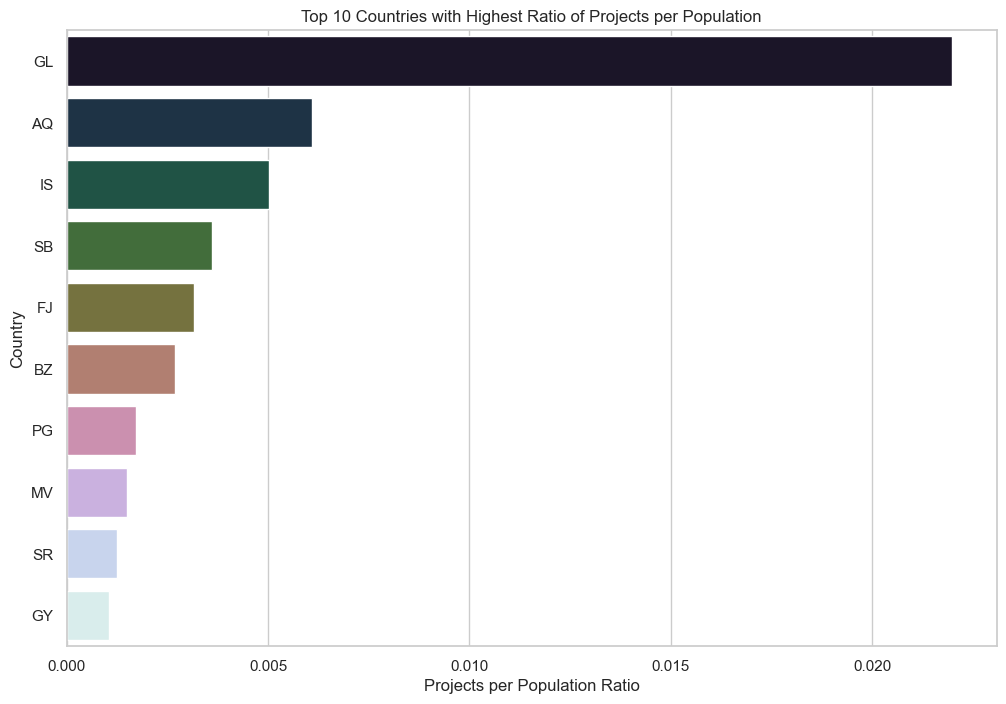
**Figure 5.** Top 10 entries in country with the most number of projects originating from them.

We know that Kickstarter is a US-based crowdfunding platform, and so it is not surprising that most projects originate from the US. It is interesting to note, however, the diversity of the origins where crowdfunding projects come from. We see projects originating from countries of different continents like Australia and Papua New Guinea (AU, PG; Oceania), Brazil (BR; South America), and China, Indonesia, and India (CN, ID, IN; Asia).

**Data Preprocessing**

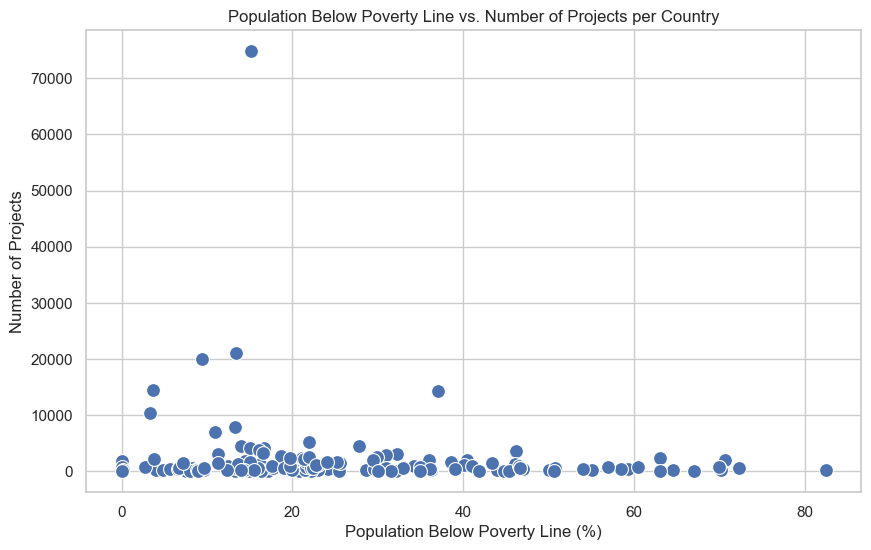
*Initial Dataset Exploration*

We display the information we obtained from comparing some of the attributes in our dataset:

****

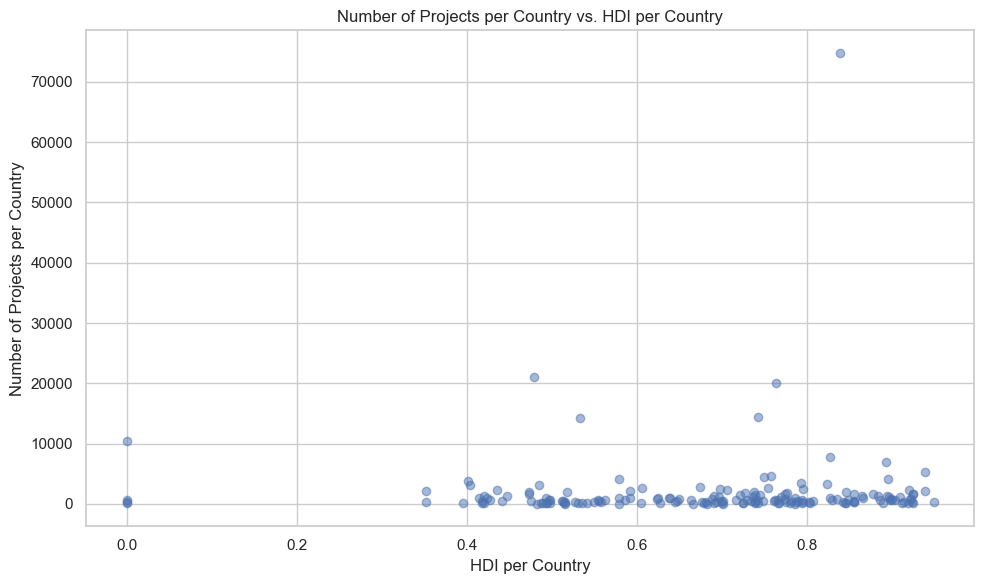
**Figure 6.** Top 10 entries in country with the highest ratio of projects per population.

From the graph, Greenland (GL) has the highest ratio of projects per population among all countries. Despite its relatively small population, Greenland's ratio indicates a remarkably active participation in crowdfunding projects compared to its population size. Additionally, notable countries such as Iceland, the Solomon Islands, and Fiji are highlighted for their similarly high ratios of projects per population.



**Figure 8.** Scatter plot of the population below poverty line vs. number of projects per country.

Visually, it appears that there may be no correlation between the poverty incidence rate of a country with the number of crowdfunding projects associated with that country. Most of the countries, regardless of their poverty incidence rate, have their total projects ranging from the 0 – 10 000 count. There are some outliers, however, like that of the US which has under 20% poverty incidence rate but has more than 70 000 projects originating from there.



**Figure 9.** Scatter plot of the number of projects per country vs. HDI per country.

A similar trend can be observed in Figure 9. Here, higher HDI may not necessarily indicate a higher number of projects created in a certain country. The outliers in this graph are the same ones as in Figure 8. One explanation for this is that poverty incidence rate is inversely related to the HDI.

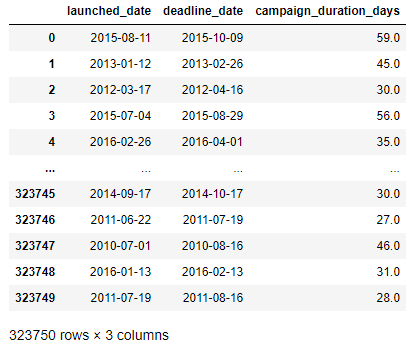
*Data Cleaning*

In this section, we outline some steps we did for cleaning the dataset.

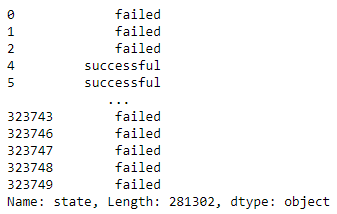
1. **Eliminated the rows with missing or NaN values from our dataset.**
   * This contains only the rows with complete data, ensuring that the subsequent analysis or modeling tasks are performed on a clean and complete dataset, free from any missing values that could potentially affect the accuracy or reliability of the results.
   * The number of rows is now down to 314 542, from 323 750.



1. **Added a new feature campaign\_duration\_days.**
   * This performs datetime manipulation by converting date columns to datetime format, calculating campaign duration by subtracting the values in launched\_date from the values in deadline\_date, and adding a new column for the duration in days.



1. **Retained the values ‘successful’ and ‘failed’ in state.**
   * This filters our dataset to include only rows where the state column contains either ‘successful’ or ‘failed’ values, discarding any rows with other states.
   * From this, the number of rows is now reduced to 281 302.



**Challenges, Assumptions, and Next Steps**

*Challenges*

* **Ensuring that the remaining rows have complete information for each column.** This is most evident in the sporadic nature of errors in the row entries, with varying mismatches on the columns. We will do further checking on the completeness of our dataset so that there would be no problems when we start with the modeling phase of the project.
* **Further integration of economic profiles in the following analyses.** As it currently stands, the economic profiles that we joined with the original Kickstarter dataset remain disconnected in the sense that most of the modeling processes we have in mind right now are focused on the original metrics. The only link they have, as of now, is the project’s country of origin.

*Assumptions*

* **Homogeneity of data sources:** All of the data incorporated in the dataset are factual, relevant, and thus underwent consistent data collection standards and methodologies.
* **Temporal, unit consistency:** All of the metrics that have a time dimension and/or measurement dimension are collected at consistent time intervals, aligned with the time of crowdfunding projects, and are represented in proper units across the entirety of the dataset.
* **Relevance of economic indicators:** While crowdfunding projects are usually more personal in nature, it is assumed that larger-context factors like economics may play a factor and have discernible impact in the success/failure of the said project.

*Next Steps*

* **Further data cleaning.** Already mentioned in the earlier bullet point.
* **Integration of text data.** We will look into the feasibility of incorporating text data for each crowdfunding campaign in the dataset. This means adding a potential column for the actual project description and relevant text markers like hashtags and the like. Having text data widens our modeling choices (NLP-based models), and may provide deeper insights in terms of influences on the success/failure of crowdfunding projects.
* **Training of the models.** So far, we are considering the following models: L1 regularization (lasso regression), Random forest classifier, and LightGBM model. After ensuring the data is cleaned and ready for modeling, we will start training our models to identify which can give better and wider insights.

**References**

Beluga. (2018, March 28). *Additional kiva snapshot*. Kaggle.<https://www.kaggle.com/datasets/gaborfodor/additional-kiva-snapshot>.

*Kickstarter Stats—Kickstarter*. (2024, March 22).<https://www.kickstarter.com/help/stats>.

Mouillé, M. (2018, February 8). *Kickstarter Projects*. Kaggle.<https://www.kaggle.com/datasets/kemical/kickstarter-projects>.