KICKSTARTER VS COVID

Jarrad Down (46643151), Jihita Sri Koya (46634766),
Joseph Menesch (43939862), Lalith Badiger (46557829)

DATA7001 Project Report
Due Date: 6 November 2020

We give consent for this to be used as a teaching resource.

Table of Contents

1.0 Executive Summary

KickStarter, despite COVID-19 and its impact on the global economy appears to be thriving in the face of adversity. This is illustrated not just through the success rates being unhindered pre-COVID and post-COVID, but also through the number of campaigns launched remaining consistent. After the 13th of March this year, when the US declared COVID-19 a health crisis, KickStarter did not slow and many interesting insights are still able to be gained.

We discovered through our project that there are varying key predictors which can impact the success of a campaign. These include category, day of release, campaign length, goal, and reward tiers. Reward tiers was another crucial aspect to determining success as the actual amounts that customers are able to pledge plays a role in campaigns meeting their goal. These reward tiers, explored through variance measures such as min, max, quartiles, and mean, were integral to predicting success alongside campaign length and the number of reward tiers present.

Different attempts were made to utilize these predictors in selecting an applicable and significant model. Methods such as K-Means were deemed to be invalid due to a lack of meaningful information provided. Other methods such as KNN Classification and Logistic Regression were discovered to be a step in the right direction for building a predictive model for success. Ultimately however, the group landed on Random Forests as the most effective model to predict success in a project and provide insight into what predictors contribute to this the most.

These random forests allowed us to discover the order of importance of predictors for each period (pre/post COVID), which gave insights into the effect of COVID on Kickstarter by comparing importance measures. Although the post-COVID model was less predictive, the predictive power of both of these models was statistically significant, even accounting for the class imbalance in the data's success measure classes. In particular, goal amount, number of reward tiers, and campaign length were found to be significant factors in a campaign's success chance, with goal importance overtaking the number of tiers importance post-COVID.

As a final note, the group highlighted some possible future extensions which could enrich the project and further aid campaigners on their road to success. Marketing materials likely play a large part in how consumers perceive campaigns and the most analyzable aspect of these materials is colour; this project could be extended to consider the colour (for example: 'warm' vs. 'cool' colours) as a predictor for success. Secondly, KickStarter is not the only crowd-funding website. IndieGoGo is another large and successful crowd-funding platform and comparative models between these two services can be used to enrich this project further. Finally, as COVID-19 is an ongoing situation, more current data will enrich the predictive models that focus on the COVID-19 period that the US currently presides in. As the pandemic worsens and continues to threaten employment and expendable income, KickStarter and other crowd-funding services may begin to be affected.

1.1 Change of Scope from Project Pitch

The project differed from the original pitch submitted in a few ways. This change was caused by the data's availability and capabilities. Due to COVID-19 being a relatively new phenomenon compared to the long history Kickstarter has accrued, time-series models were not considered beyond exploratory data analysis. The reason for this is the models we chose to implement to predict success were not effective when considering time. The original pitch stated that this project would track the change in predictors relative to COVID-19 presence worldwide. Instead, the project scope was narrowed to focus on only the US. Furthermore, the project did not track the change in predictors relative to time or COVID-19 presence, but rather focused on the effectiveness of predictive models before the US declared COVID-19 a crisis (13th March 2020) and after this date. In addition to this, we aimed to curate data from early 2019 to as late as we could in 2020 to gain a balanced perspective of before this date and after.

1.2 Related Works

In our initial research around the topic of Kickstarter, we discovered several sources of data on Kaggle and associated projects. These projects we discovered however were mostly around cleaning data and exploring the data, no predictive models were explored, and no element of COVID-19 was present in these projects as they did not curate Kickstarter data in 2020, unlike WebRobots, the source of our data. While the data from Kaggle would be applicable for part of our project, we required recent data which Kaggle was unable to provide; data only promised dates around 2016 and 2018. Appendix A illustrates examples of what these projects generally explore.

2.0 Data Science Process

This project strictly followed the generic Data Science project explored throughout this course. Using Design Thinking, the group formulated a human-centric problem concerned with the potential effects of COVID-19 on the ability of campaigns of Kickstarter. After formulating this problem, data was sought to explore this problem best with key aims in sight; mainly the recency of data was the primary concern. After securing the data, the limitations and features of the data were explored. This stage highlighted the need for cleaning and further data gathering the form of web-scraping. Once the data was suitable for use, the group applied several modelling techniques to draw potential out of the data in whatever form was optimal. Once these key steps are complete, a story concerned with the journey to maximizing success was constructed to highlight the value and insights of this project.

2.1 Problem Solving with Data

KickStarter has been in operation since 2009 and the market of crowd-funding was revolutionized by it. Today, it can be difficult to determine the most effective methods to increase the likelihood of success. KickStarter does provide advice to campaigners, however it is ultimately a contextual issue as no two campaigns were created equal. The merit of ideas, products and innovations may largely contribute to success; however, this merit is too semantic and contextual to be measured or analyzed effectively. Instead, this project aims to affect what change is possible through the ideal selection of parameters which may heighten the odds of success.

Campaigners are the key stakeholder of this project as we aim to maximize their odds of a successful campaign during COVID-19; this was achieved in two methods. Firstly, through highlighting key decisions that a campaigner can make, and secondly by providing a model through which campaigners could predict the success of their campaign relative to what predictors are available. Success in this instance as defined where a project meets is allotted goal, irrespective of whether the campaign is cancelled for other reasons. The reason for this distinction is to clarify that this project aims to maximize projects meeting their goal, not whether they decide to continue with the project irrespective of meeting their goal or not.

A scenario in which our project may be used is if a campaigner is planning the launch of their campaign for a Table Top Game related product. Through our project, we can advise whether launching this product is a good idea based on the category of product and recommend the length of the campaign and perhaps even what day of the week to launch. After which the campaign manager can provide a predictive model with their proposed reward tiers alongside the aforementioned information and can estimate their chances of success *solely based on the information provided.* This prediction is observed through both pre-COVID and post-COVID models in order to determine the change in predictability of success.

2.2 Getting the Data I Need

Kaggle was one of our candidates in the pursuit of Kickstarter data. But the data available in Kaggle was only for the years 2018 and prior. The data requirements for this project include considerable amount of data before and after the COVID outbreak. In narrowing the scope of this project, the refined data requirement is concentrated on data just before COVID and after the outbreak. Kickstarter data for the year 2020 is also important in analyzing the impact of COVID on predictability. Therefore, the KickStarter data available in Kaggle will not serve the purpose.

WebRobots provided the solution for our data requirement through the URL in Appendix I. WebRobots curated this data by web scraping KickStarter through each category. The data relates to all the Kickstarter projects up until the date of collection by web scraping the Kickstarter website. This process is repeated once in every month and the scrape data is made available on the WebRobots website. The data is stored in multiple csv files and will be available for use.

The Kickstarter data gathered from WebRobots contains a large variety of data related to the Kickstarter projects. They contain information related to important dates of a project such as the day the project is created, the date on which the project is launched, and the deadline of the project deadline. Apart from these dates, the data also contains semantic data of the project such as project profile, creator profile and the location details. The performance indicators of a project such as pledged amount, backers and goal are also available.

Beyond this, the data contained useful conversation rates that would be necessary when exploring relative success across the world in our initial exploratory data analysis. The data also contained several other less useful columns, such as URLs to the page of the KickStarter, URLs to the display pictures of the campaign, URLs to the users curating the campaign.

2.3 Is My Data Fit for Use

This section discusses the exploration of the data in determining whether it was viable and applicable to our project, and the data quality improvements made to optimize its usefulness.

2.3.1 Additional Data Gathering

A primary feature of Kickstarter is the ability for backers to allocate their contribution towards a certain reward tier, as defined by the campaign runner, which promises certain campaign-related rewards if the campaign is successful. For example, a campaign for the development of a graphic novel might have a reward tier which gives a paperback version of the novel to the contributor, and a higher-cost reward tier which gives a hardback version. We hypothesized that such information might have an effect on a campaign's success chance, as the rewards likely influence the contributor's decisions. As the Kickstarter dataset did not include this reward information, we decided to develop an algorithm to web scrape this data for each campaign in our cleaned dataset. At the time of running the algorithm, this was about 36,000 campaigns. Kickstarter's copyright licensing permits the use of all data on their website for non-commercial use, so there's no legal issues with this process.

The web scraping was done in Python 3.8, using several non-standard libraries (Appendix B). Firstly, the algorithm extracted the rewards URL for each campaign, and put them in a pandas data frame along with the corresponding campaign's id. The URL was then loaded, and the html code was parsed through a html parser class. This class did a search for the relevant tags (found by inspection) and saved the information within those tags. This saved the reward cost for each tier, along with the corresponding number of backers that tier had (see reward_webscraper.py). A sample of this code is given below. Due to the time the algorithm would take, the list of reward information was appended to a CSV file after every 5 campaigns were web scraped, so as to avoid losing progress if the script was interrupted. Overall, this algorithm had a runtime of about 22 hours. After cleaning the reward data (removing currency symbols, removing strings, vectorizing – see reward_output_cleaner.py), this data was ingested into the main Kickstarter dataset. This was done using an inner join on the id column.

```
def find_rewards(url):
    f = urllib.request.urlopen(url)
    text = f.read()
    f.close()
    parser = LinkParser()
    parser.feed(str(text))
    return parser.get_rewards()
```

Code figure 1: Snippet of Web-scraping Code

2.3.2 Data Cleaning

The data collected from WebRobots has details of all the projects from 2016 until the scraping date. Several measures of cleaning by filter were implemented to best fit the scope of the project. The data is filtered for all the projects created after the 1st of January 2019, to get roughly equal time periods before and after the COVID outbreak. Additionally, data is filtered for all the projects in USA as most of the projects are from USA and also to ensure consistency in the analysis. Duplicate records identified in the raw data from WebRobots are removed with project id as the identifier. This cleaned data is enriched with the additional data gathered as a part of web scraping at the project instance level.

This additionally obtained data of web-scraped reward tiers was converted into a JSON format. For convenience, these JSON fields were transformed into a comma separated string of reward tiers for ease. A final scan of the data showed there were large and unnecessary columns that needlessly clogged the data. These columns were trimmed for ease of use as they played no part in our project.

A sample of the original, uncleaned data can be found in the file ‘Uncleaned_KS_Data.csv’ and the cleaned data used in exploratory analysis and modelling can be found in ‘KS_Data_Final.csv’. The code used to perform this cleaning is contained in the R files labelled ‘Data_cleaning_v1’ through ‘v4’.

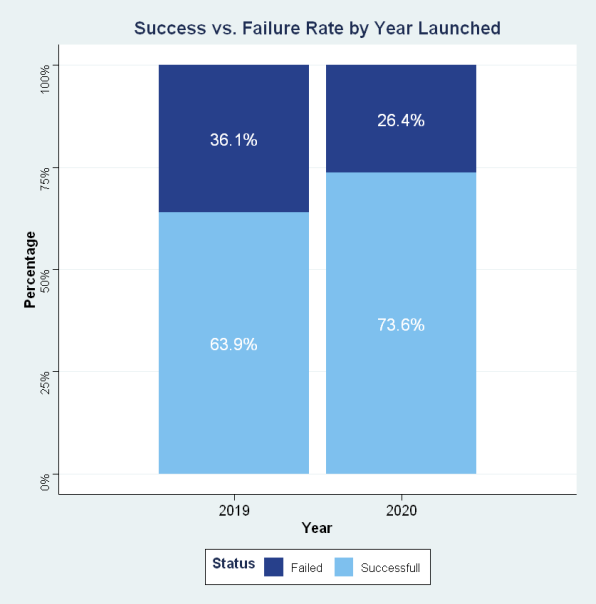
2.3.3 Data Imputation

Data imputation was not a process that needed to be considered for our project due to the state of our data on collection. While there were (very few) missing values, the columns containing any missing values comprise of semantic information incapable of being imputed. Furthermore, this small amount of missing data seemed to be missing completely at random as a result of web scraping errors, so these columns were simply excluded when necessary.

2.3.4 Exploring the Data (EDA)

This section illuminates the data through exploratory data analysis and the identifies features that the we were able to extract and consider for model selection.

2.3.4.1 Project Status Analysis



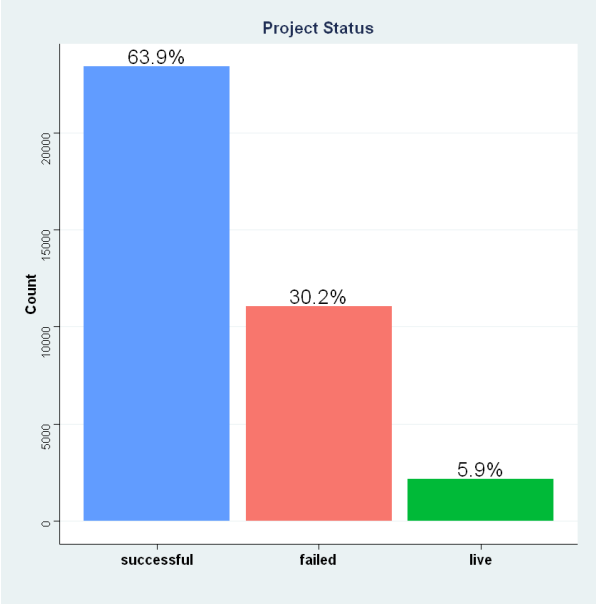


Figure 1: Status Distribution by Year and by Status

Our dataset contained data ranging from January of 2019 to July of 2020. And, from the first plot, we can see that the **success rate in 2020 has increased by around 10%** as compared to 2019, from 63.9% to 73.6%.

Overall, 63.9% of the projects were found to be successful, 30.2% failed and 5.9% which are still “live”. Hence, **more projects are successful than failed**. "Cancelled" and “suspended” projects are removed from our analysis here.

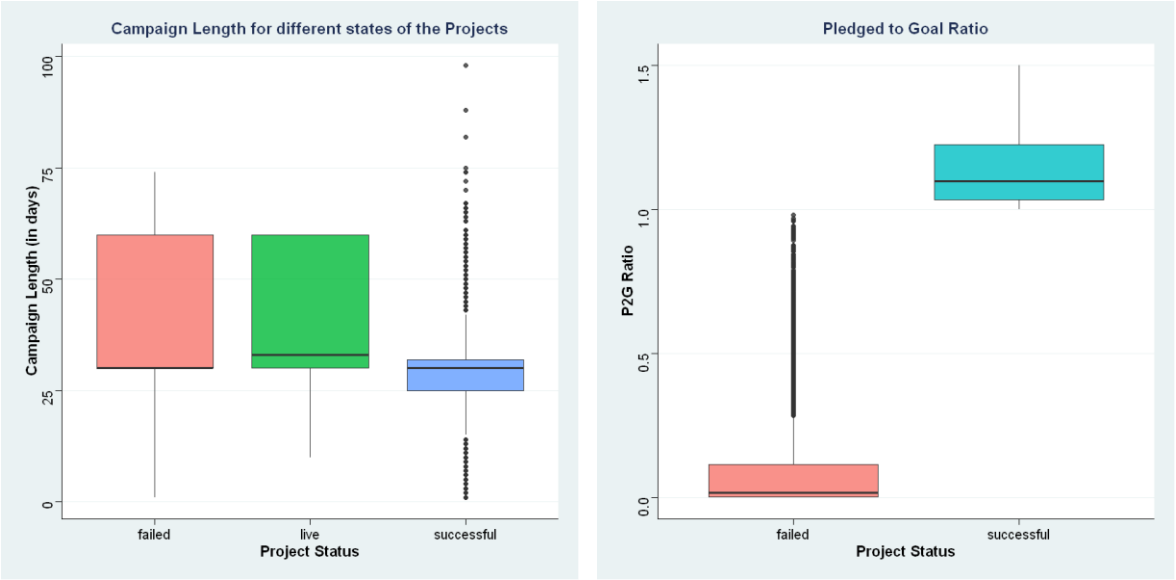


Figure 2: Box Plots of Project Status for Campaign Length and Ratio of Pledged to Goal Amount

As seen from the first boxplot, successful projects have a lot of outliers, but the size of the box is small indicating a **high concentration at around 30 days which makes it the ideal campaign length for successful projects**. The upper quartile for failed and live projects is large, indicating that there is a right skewness in the distribution after 30 days.

From the second boxplot, we can see that typically **failed projects won't even reach 20% of its required goal but successful projects get 5% to 25% excess amount pledged over the set goal**.

2.3.4.2 What types of projects are most popular?

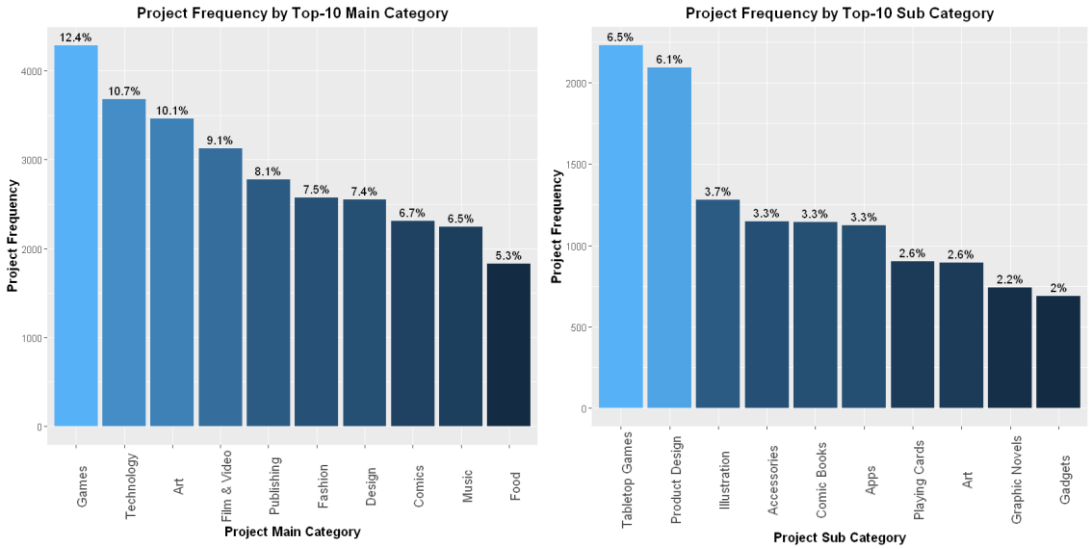


Figure 3: Project Distribution by Main Categories and Subcategories

After filtering out null values from our dataset, **the most popular main categories by count are found to be Games, Technology and Art** at 12.4%, 10.7% and 10.1% respectively. The most popular sub-categories are found to be Tabletop games (like board games, dice games, etc.) and Product design (like modular shelf, custom designed backpack, etc.) at 6.5% and 6.1% respectively.

2.3.4.3 What types of projects are being funded?

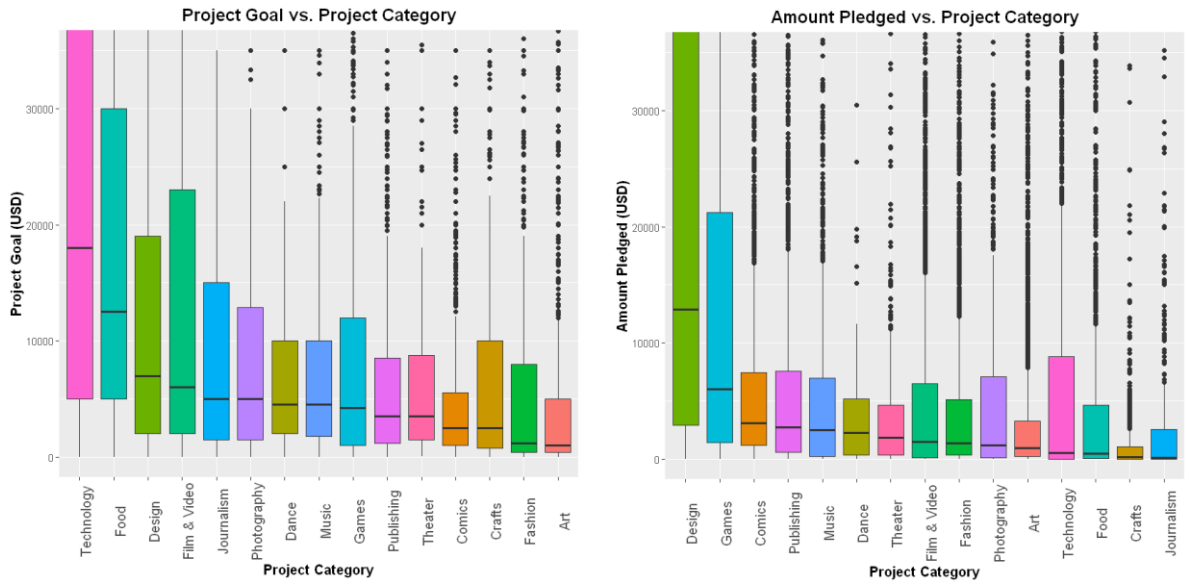


Figure 4: Goal and Pledged Amounts by Project Category

The above boxplots show the distribution of project goal and the amount pledged with its respective category arranged in descending order of their medians.

When we see the Project goal in Technology, the median is very high at around $17,000 USD but when we see the same category in the Amount pledged plot, we can see that the pledged amount is really low. Even though there are many outliers, from the third plot we can see that the success rate of Technology category is among the lowest i.e. 31.6%.

Similarly, when we see the design category, the pledged amount is very high since the goal is kept relatively low. And hence, the success rate is highest for the design category, at 89.1%.

From these plots, it can be hypothesized that **lower goal leads to a higher amount pledged and hence there is a high chance of the project being successful.**

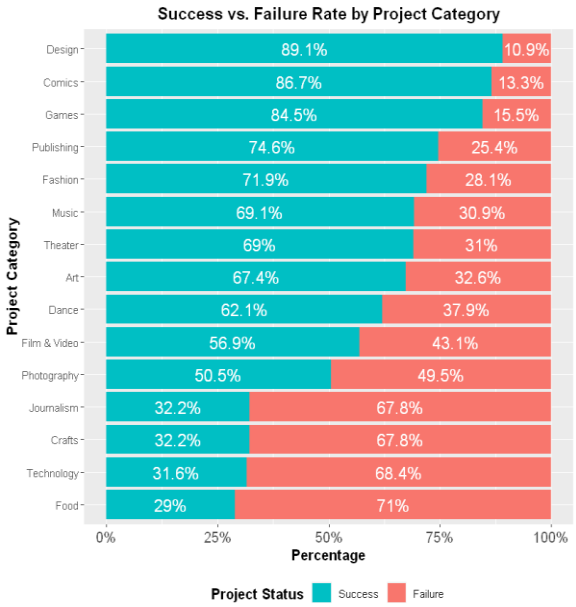


Figure 5: Success and Failure Rates by Category

2.3.4.4 Does campaign length impact success rate of the project?

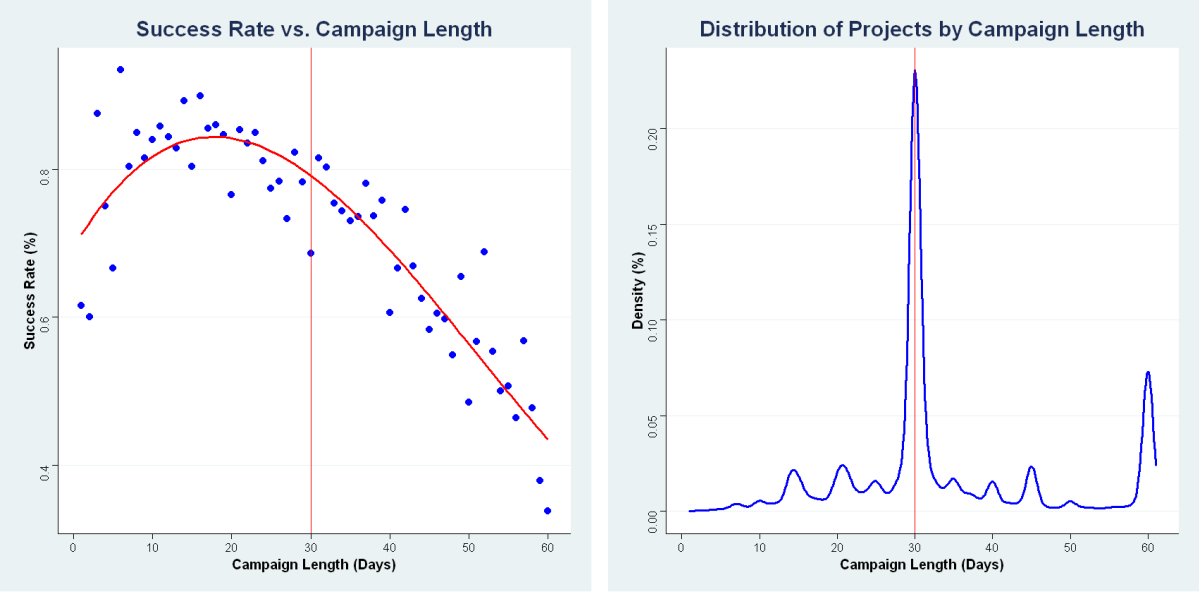


Figure 6: Success Rate and Distribution of Campaign Length

Kickstarter was founded in 2009 and until 2011, the maximum length of the project campaigns was 90 days. But since 2011, the max campaign length is capped at 60 days. Kickstarter has previously advised that "projects on Kickstarter can last anywhere from 1 - 60 days. We've done some research and found that projects lasting any longer are rarely successful. We recommend setting your campaign at 30 days or less." (Kickstarter, 2020) This quote from the company is directly aligned with the data, and their advice has been noticed by campaigners.

In order to test this claim, we plotted campaign length vs success rate with a non-linear regression of 3rd order polynomial and found out that the success rate during the initial few days is higher. **After 30 days the probability of the project reaching its goal is significantly reduced.**

From the second plot, we can see a **peak at the 30th day** which is the recommended length by Kickstarter. Our theory is that having more time does not create a sense of urgency. Instead, it makes it easier for backers to procrastinate, and sometimes they forget to come back at all. When a project launches the creator's most avid fans rush to show their support and as time runs out, people who have been sitting on the side-lines are motivated to finally take action. And hence, **choosing a shorter duration better positions a project for success.**

2.3.4.5 Impact of COVID-19 on Kickstarter Projects

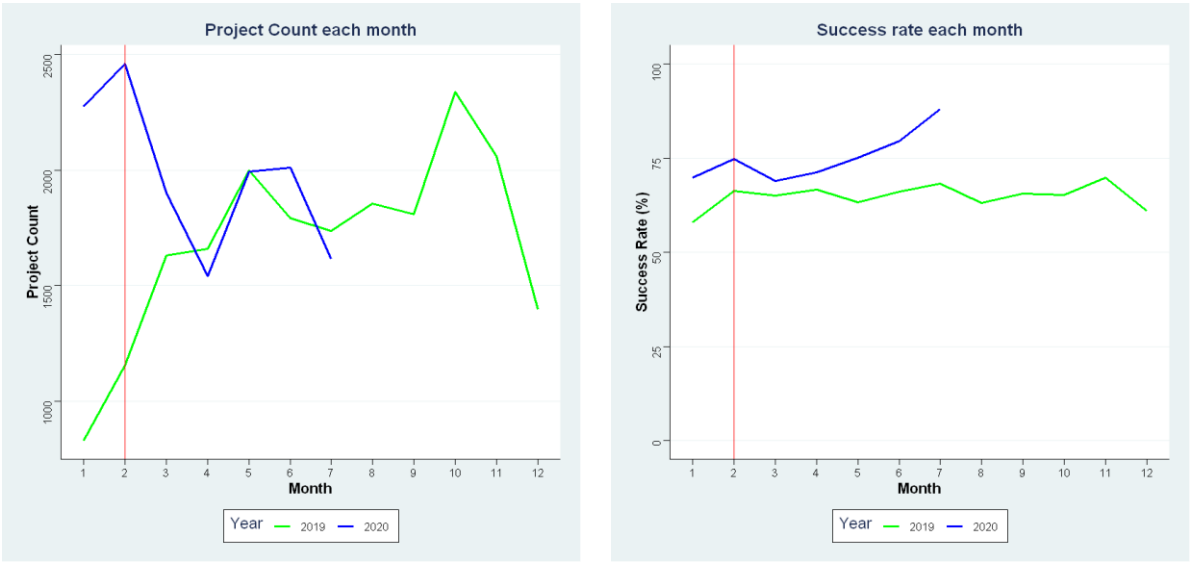


Figure 7: Project Count and Success Rate by Month across 2019 and 2020.

Since the US was highly affected by COVID and due to Kickstarter being highly used in the US as compared to other countries, we just considered US data for this analysis. The blue line represents 2020, and the green one represents 2019. Considering the pandemic started in February of 2020, represented by red x-intercept line, from the line plots we can see that 2020 started off really well

in terms of project count and success rate. As compared to last year's statistics, **2020 is at par or even better than 2019 in terms of project count and success rate**.

2.3.4.6 Does launch weekday affect success?

During EDA, we also examined some less obvious features, including various feature transformations and how they influence success. One such feature was the weekday on which the project was launched, which was determined by the given campaign launch date. As shown in the figure below, there seems to be a large number of projects launched on Tuesdays, and significantly fewer projects launched on weekends. Furthermore, the ratio between successes and failures seems to the largest on Tuesday, indicating a higher chance of campaign success when launched on Tuesdays.

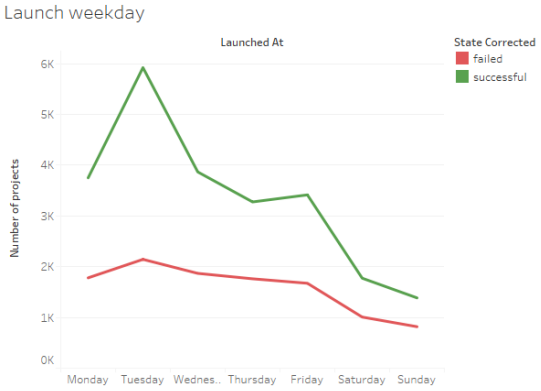


Figure 7b: Project count by launch day

2.4 Making the Data Confess

In order to determine potential predictive models, we formulated a list of success measures (and their variable types), and a list of independent variables (and their variable types). Following EDA, we narrowed down this list to a subset of variable combinations and determined the best modelling procedures to use for each combination based on the data types and hypotheses. Based on EDA, the most promising success measure to use was defining successful campaigns as those which had a pledge amount greater than or equal to its goal and defining failures as all others. This includes defining a suspended/cancelled campaign as successful if it reached its goal. From the models we developed, some were invalid, some were valid but had sub-optimal performance, and one proved to have quite good predictive power. Each model was applied only to US campaigns before and after the declaration of the COVID pandemic by the united states (13th March 2020). These models are described in detail below. Note that these models were developed in R using the packages listed in Appendix C, with the help of some additional resources. (Bali, Sarkar, 2016; Kelleher et. al., 2015; Chakravarti et. al., 1967)

2.4.1 Invalid Modelling

2.4.1.1 K-Means Clustering

To start with, we wanted to investigate any clusters of contributor's behavior corresponding to different categories, and whether it's changed since COVID. We measured contributor behavior by 2 dimensions: Number of backers and amount pledged for each campaign. Using this metric, we could see, for example, whether some categories attracted many backers with low pledge amounts, and vice versa. After normalizing the log of the predictors, k-means clustering was used with 10 clusters. The results of this clustering are shown in Appendix D, along with a mosaic plot for the categories in each cluster. This model produced no clear clusters outside of the null hypothesis regression (H0: Number of backers is proportional to amount pledged for each category), so no insights can be drawn from this model which aren't already clear from analysis of category success. The code for this modelling is available in 'kmeans_backers_pledged.R'.

2.4.2 Sub-Optimal Modelling

2.4.2.1 KNN Classification

After considering invalid or inapplicable modelling, the team began discovering some useful, yet sub-optimal predictive models such as K Nearest-Neighbor classification. The classification was performed over two factors: successful and failed. The model attempts to classify success based on reward tier variance measures (minimum, maximum, quartiles and mean) relative to the total size of the goal.

The training process for the classification model automatically determined the ideal value of K, where accuracy is maximized. This was observed to have been somewhere between 23 and 33, as shown in Appendix J. After the optimal K was determined, the model was trained and tested and was found to be slightly more accurate than guessing. As shown in Appendix K, the accuracy was not ideal, with the sensitivity and specificity not within ideal ranges, both below 0.66, with sensitivity being quite low at approximately 0.57. This was, however, a step in the right direction as we began to move closer to building a model that could make reasonable predictions. Even if this model were more accurate, the issue of the model being very uninterpretable will prevent the predictions from being too meaningful, as we have no ideas concerning what predictors are important and which do not contribute largely to success. Learning from this model, the project advanced into more applicable models centered around predicting success and failure as factors rather than as a relative measure of success. The code for KNN classification can be found in the file 'knn_classification_complete.R'.

2.4.2.2 Logistic Regression

Feature analysis during EDA (Appendix E) told us that the number of reward tiers was the greatest predictor of success. To validate this hypothesis, we built a univariate logistic regression model which determines success based solely on the number of reward tiers and tested it on a randomly sampled 30% test set (not used for training), available in 'logreg_numtiers_state.R'. The pre-COVID model is illustrated in Figure 8, along with prediction metrics in Table 1. The post-COVID model and predictive metrics are given in Appendix F. For pre-COVID, the accuracy was greater than a Bayes optimal classifier (i.e., naïve guessing) with statistical significance ($P < 2.2e-16$), but this was not the case for the post-COVID model ($P=0.11$). Pre-COVID gave $B0 = -0.66$ and $B1 = 0.18$, whereas post-COVID gave $B0 = -0.47$ and $B1 = 0.23$, implying that the number of tiers increases success chance.

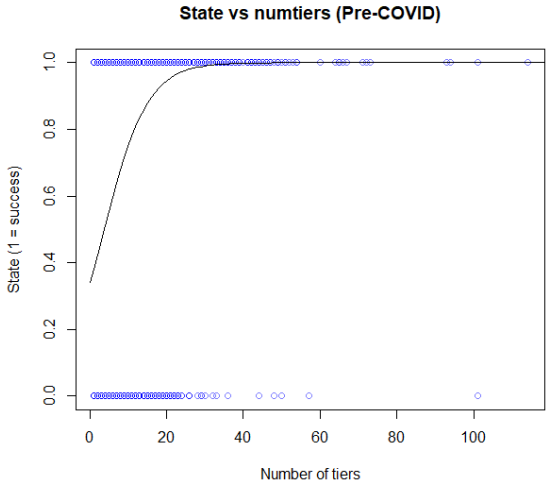


Figure 8: Logistic regression plot for US Pre-COVID

Pre-COVID		*Actual*	
		Failed	Successful
Prediction	Failed	583	267
	Successful	973	2602

Performance metric	*Value*
Accuracy	0.7198
95% confidence interval	(0.7063, 0.733)
No information rate	0.6484
P-value (Acc > NIR)	< 2.2e-16
Sensitivity	0.9069
Specificity	0.3747
Positive predictive value	0.7278
Negative predictive value	0.6859
Balanced Accuracy	0.6408

Table 1: Logistic regression performance for US Pre-COVID

Firstly, as shown in the figure, there is a class imbalance towards successes, which is also the case for post-COVID models. Assuming the actual distribution is similar to that the model was trained on, which we have no reason to believe otherwise, this shouldn't be a problem for this model's predictive power. Secondly, it is clear that the campaigns with more reward tiers tend to be successful. However, it isn't clear whether this is due to the class imbalance. Even if the number of tiers is drawn from the same distribution for both successful and failed campaigns, the class with the higher number of cases (successes in this case) would tend to have more campaigns at the tail ends of this distribution. To test whether these distributions were the same, we used the Kolmogorov-Smirnov test, available in kolmogorov-smirnov_test.R, which measures the "distance" between the observed ECDFs of each class (Chakravarti, et. al.). This test returned D=0.35385, with a P-value < 2.2e-16 pre-COVID (and similar post-COVID), allowing us to reject the null hypothesis that the distributions are the same. Therefore, we can conclude that the number of tiers increases success chance pre-COVID, as predicted by the logistic regression model.

2.4.3 Random Forests Predictive Model

Finally, we used random forests to model success based on the most relevant predictors given in feature analysis. This included goal, campaign length, number of reward tiers, and reward cost quartiles. Each of these variables were normalized to 0 mean and 1 standard deviation, the data was randomly sampled into a 70%/30% training/test split. Furthermore, the random forests' hyperparameters, namely the number of trees and the node size, were tuned with the e1071 library, using 10-fold cross validation, illustrated in the code output snippet below (Code figure 2). As the number of training examples post-COVID was only 2167, cross-validation was useful here. The random forest was trained and tested, and performance metrics were evaluated, given in appendix G. Due to the class imbalance, recall-precision curves and positive/negative predictive values were used as the primary performance measures, rather than the ROC curve and its AUC and sensitivity/specificity. The post-COVID recall-precision curve is illustrated in figure 9.

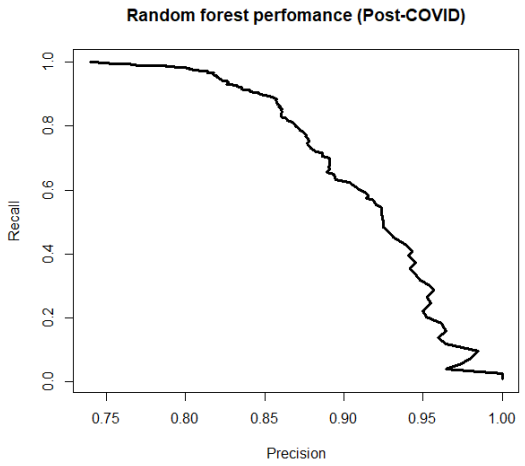


Figure 9: Random Forest PR Curve

```
Parameter tuning of 'randomForest':

- sampling method: 10-fold cross validation

- best parameters:
 nodesize mtry ntree
        5    3  2000

- best performance: 0.1624253
```

Code figure 2: Random forest tuning

Both models gave statistically significant prediction accuracy, at 0.75 and 0.70 balanced accuracy (P-value < 2e-16 and P-value = 1.09e-6) for pre-COVID and post-COVID respectively. Figure 11 shows the variable importance for both models, as measured by mean decrease in accuracy (MDA). This information is also tabulated in Appendix H. Note the difference in the axis's ranges due to the extra difficulty of predicting success/failure post-COVID. By examining figure 10, we can see that, while the top three predictors remained the same across both periods, the goal amount became the top predictor post-COVID, overtaking the number of tiers. The top three predictors were far more influential than the specific reward tier cost amounts, with the minimum tier cost having significantly less predictive power than the other quartiles. Furthermore, we know from the model that the most influential predictors towards success in particular are goal and number of tiers (given by success importance, in Appendix H), and choosing an appropriate campaign length would help to deter failure (failure importance, in Appendix H). The code used for random forest modelling is available in 'random_forests.R'.

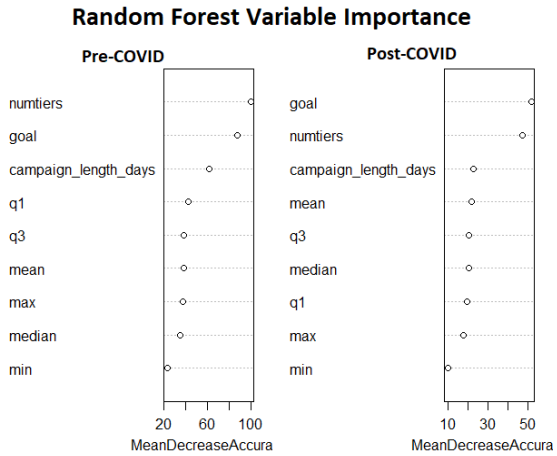


Figure 10: Random Forest Variable Importance Plot

2.5 Storytelling with Data

Our initial project goals were aligned with storytelling regarding the differences between pre-COVID Kickstarter campaign success predictors and how they have been manipulated by the change in economic climate brought on by the pandemic. Due to the previous reasons discussed that caused the change in scope of our plans, the story best told through this project is a roadmap to success. This roadmap was built on a multi-faceted consideration of all contributing parameters to success. These characteristics include day of the week of launch, category of the campaign, length of the campaign in days and a plethora of measures (such as mean, minimum and maximum) of the available reward tiers for supporting a campaign.

We discovered through EDA that the best days to launch your campaign were early in the week. The difference, as shown in the exploratory analysis, was staggering enough to contribute to our roadmap to success. Furthermore, as per exploratory analysis and advice from KickStarter themselves, the advised campaign length is 30 days.

Beyond these simple predictors of success were more complex predictors, namely in the reward tiers. Through careful examination of the reward tiers in number, mean and general distribution including the upper and lower bounds as well as quartiles, we were able to create an effective predictive model. This predictive model, alongside the advice we are able to provide campaigners present a pathway to success for anyone with a product, innovation or idea they want to market to everyday people. This story was presented in the presentation for this project and was well received, the audience had a key understanding of how success could be influenced through manipulating these measures.

In telling the story throughout the entire presentation for this project, a variety of techniques were used. Change over time was demonstrated to illustrate that the trend of success on KickStarter was unaffected through the influence of COVID-19 even as it worsened from April onwards. The other main technique was the drill-down, as we constantly re-established the context in a deep-dive of what predictors were important out of all possible parameters used to establish a campaign. We demonstrated through explanatory and statistical measures why these predictors are important and even how they compare to each other as we highlighted which predators were more important than others. However, we were extremely cautious in presenting the story and emphasized that these measures which can be implemented within a campaign are not exhaustive, nor are they a guarantee of success. As said in establishing the problem space for this project, we predict that the biggest influence of success is the inherent value or merit of the product or idea itself; we are only attempting to maximize what can be maximized in a general campaign.

3.0 Feedback Summary

Overall, our project was positively received. Some clarification surrounding the top predictors of our random forest model was needed, however the presentation of the project received no critique from the immediate audience. The project was deemed a success by peers with minor critique and insightful guidance as to how the project can be improved or deepened.

The peer feedback received mostly coincided with the audience's reaction. Feedback indicated that our project was very complete and contained all steps of the data science process and conveyed them in a clear and coherent manner. It was noted that the choice of problem space was well defined and extremely relevant to the state of the world. The aim of the project was successful according to feedback and the depth of data cleaning and filtering was commented on as well explained and effective. Another key positive raised in feedback was our exploration of differing models and explaining why some didn't work, why some were sub-optimal and the resulting choice of a best model. Additionally, the project was praised for not accepting just the data that was available but performing our own web scraping to enrich the project.

After considering peer feedback further, the group was able to identify some areas for improvement. Mainly these areas were focused on potential improvements to the presentation and alterations to modelling and how the project was conducted. There were several requests for clarification surrounding why decisions were made or how success was measured. Our presentation could have clarified why Random Forests as a model was used as opposed to our other sub-optimal models in a clearer way. The reason for this being that its predictive qualities were more significant, and it provided insight into what predictors played the largest roles, which were explained. Success as a measure felt unclear to the audience as we never clarified in the presentation that we judged success to be any campaign that reached its goal, even if the campaign was cancelled. We considered this to be a more accurate version of success, as if a cancelled campaign met its goal, it failed due to non-monetary reasons. Since our project is entirely based around maximizing the chance of monetary success of the campaign, this felt valid. The final advice we received regarding the presentation was the over-use of words on slides, in future the presentation's visuals should be succinct and direct.

The feedback received regarding modelling and data was focused around exploring these aspects more deeply. Despite clarifying that a future expansion of this project is to obtain more data as COVID-19 continues, we received feedback recommending this further data collection. There are currently two more batches of this data located on WebRobots which could be used to enrich the project if time allowed; however, as we are required to web scrape reward tiers further, this is a time-consuming process meaning the extension is not possible within the time frame provided. Another recommendation was to explore the failures of KNN Classification and Logistic Regression further; this was actionable advice that was implemented in our report as we thoroughly explain why these measures were either invalid or sub-optimal compared to Random Forests. A more specific recommendation was to use stratified sampling when performing logistic regression to improve the imbalance; the class imbalance in the logistic regression model was resolved in this report also.

4.0 References

Kickstarter. (2020). What is the maximum project duration? Retrieved from https://help.kickstarter.com/hc/en-us/articles/115005128434-What-is-the-maximum-project-duration-

WebRobots. (2020). Kickstarter Datasets – Web scraping service. Retrieved from https://webrobots.io/kickstarter-datasets/

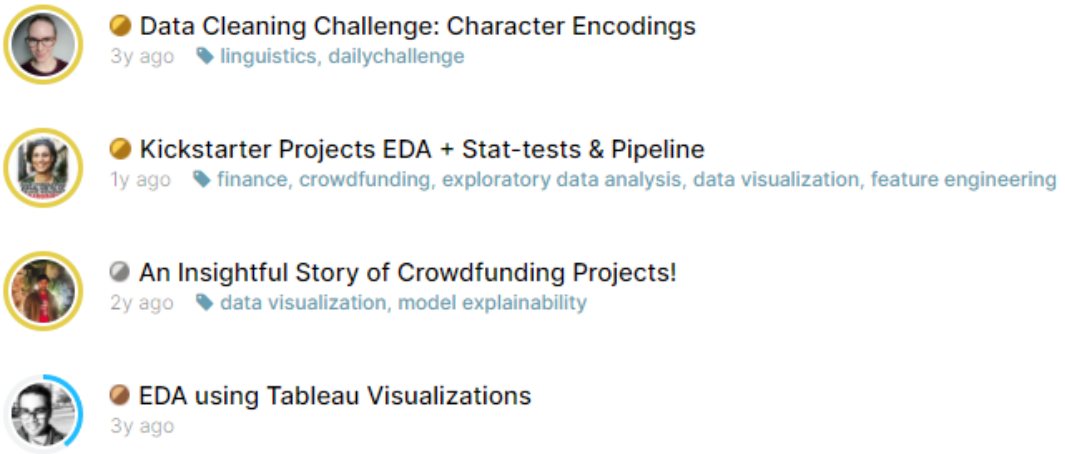
Chakravarti, Laha, and Roy. (1967). Handbook of Methods of Applied Statistics. Volume I, John Wiley and Sons, pp. 392-394.

Bali, Raghav, and Sarkar, Dipanjan. (2016). R Machine Learning By Example. Packt Publishing, Birmingham UK.

Kelleher, John D., MacNamee, Brian, and D'Arcy, Aoife. (2015). Fundamentals of machine learning for predictive data analytics. MIT Press, Cambridge USA.

5.0 Appendices

Appendix A – Examples of Related Kaggle Projects



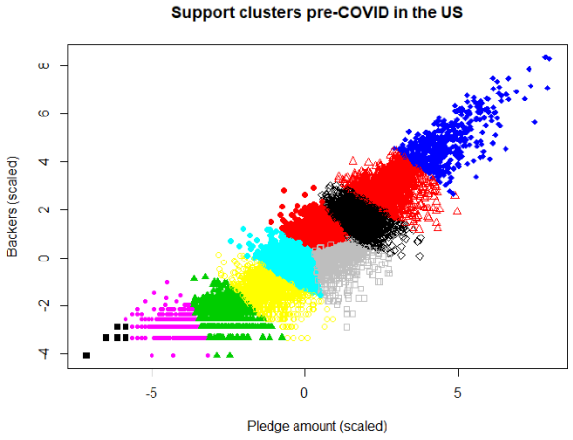
Appendix B – Python Libraries Used for Web Scraping:

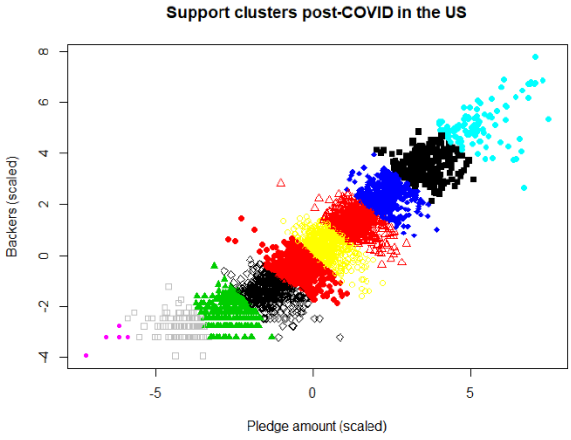
- html
- urllib
- pandas
- yqdm
- pathlib
- ast

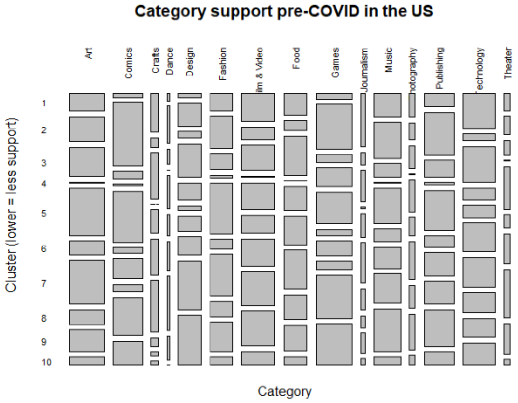
Appendix C – R Libraries Used for Modelling:

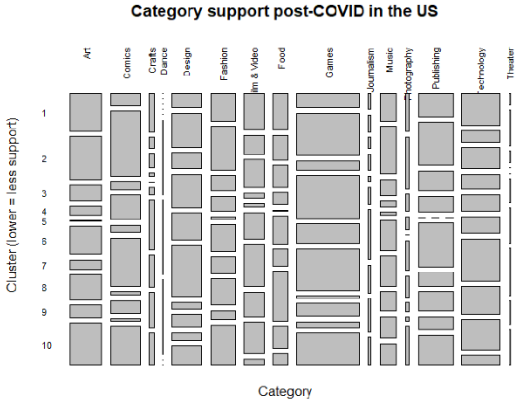
- readr
- dplyr
- tidyverse
- ROCR
- caret
- randomForest
- e1071
- pROC
- quantable
- lubridate

Appendix D – K-Means Clustering and Category Mosaic Plot









Appendix E – Feature Analysis Output (feature_analysis.R)

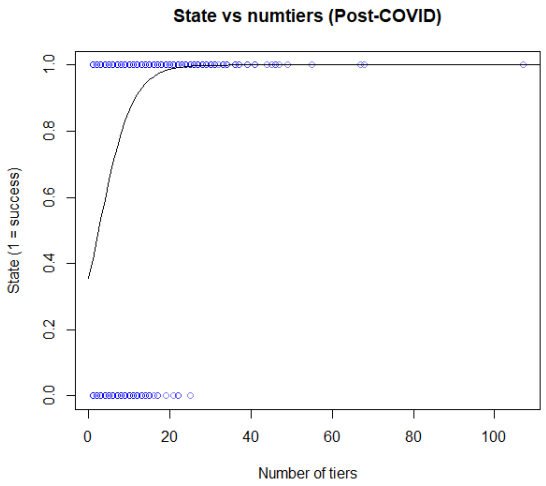
Pre-COVID (raw output)

```
Recursive feature selection
Outer resampling method: Cross-Validated (20 fold)
Resampling performance over subset size:
 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.7215 0.3111    0.01760 0.04644
         2   0.7680 0.4570    0.01625 0.04051
         3   0.7817 0.4876    0.01667 0.04149
         4   0.7651 0.4624    0.01625 0.04035
         5   0.7802 0.4946    0.01549 0.03700
         6   0.7874 0.5097    0.01770 0.04300
         7   0.7883 0.5112    0.01789 0.04168        *
         8   0.7875 0.5092    0.01799 0.04142
         9   0.7861 0.5063    0.01750 0.04075
The top 5 variables (out of 7):
   numtiers, goal, campaign_length_days, q1, q3
> varImp(rfe.results1)
                       Overall
numtiers             107.16551
goal                  82.96536
campaign_length_days  64.03093
q1                    33.54494
q3                    32.50278
max                   30.63104
mean                  30.01699
median                28.87658
```

Post-COVID (raw output)

```
Recursive feature selection
Outer resampling method: Cross-Validated (20 fold)
Resampling performance over subset size:
 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.7578 0.2250    0.02609 0.11376
         2   0.7899 0.3868    0.03222 0.09809
         3   0.8046 0.4390    0.02369 0.07243
         4   0.8055 0.4494    0.02970 0.08663
         5   0.8124 0.4681    0.02864 0.08549
         6   0.8172 0.4791    0.02783 0.08451
         7   0.8159 0.4734    0.02882 0.08714
         8   0.8202 0.4851    0.02682 0.08057
         9   0.8272 0.5052    0.03292 0.10097        *
The top 5 variables (out of 9):
   numtiers, goal, campaign_length_days, q3, max
> varImp(rfe.results1)
                       Overall
numtiers             51.576643
goal                 49.321007
campaign_length_days 22.301779
q3                   15.952854
max                  15.842481
mean                 14.413830
q1                   13.657695
median               12.930042
min                   6.249047
```

Appendix F – Logistic Regression Model and Metrics (Post-COVID, 'Numtiers' Vs State)



Post-COVID		*Actual*	
		Failed	Successful
Prediction	Failed	58	42
	Successful	170	660

Performance metric	*Value*
Accuracy	0.7720
95% confidence interval	(0.7437, 0.7986)
No information rate	0.7548
P-value (Acc > NIR)	0.1182
Sensitivity	0.9402
Specificity	0.2544
Positive predictive value	0.7952
Negative predictive value	0.5800
Balanced Accuracy	0.5973

Appendix G – Random Forest Performance Metrics Tables

Pre-COVID		*Actual*	
		Failed	Successful
Prediction	Failed	940	374
	Successful	570	2541

Performance metric	*Value*
Accuracy	0.7867
95% confidence interval	(0.7743, 0.7987)
No information rate	0.6588
P-value (Acc > NIR)	< 2e-16
Sensitivity	0.8717
Specificity	0.6225
Positive predictive value	0.8168
Negative predictive value	0.7154
Balanced Accuracy	0.7471

Post-COVID		*Actual*	
		Failed	Successful
Prediction	Failed	114	52
	Successful	128	636

Performance metric	*Value*
Accuracy	0.8065
95% confidence interval	(0.7796, 0.8314)
No information rate	0.7398
P-value (Acc > NIR)	1.090e-06
Sensitivity	0.9244
Specificity	0.4711
Positive predictive value	0.8325
Negative predictive value	0.6867
Balanced Accuracy	0.6977

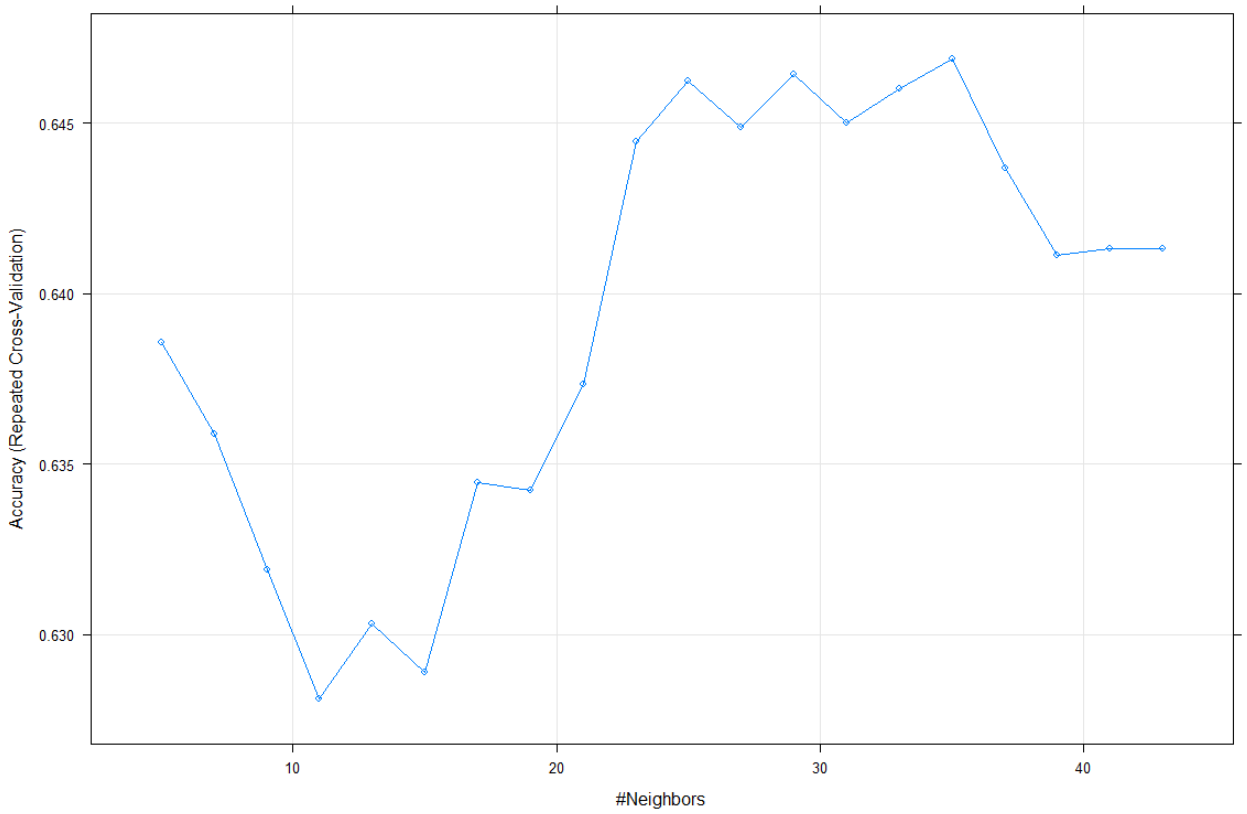
Appendix H – Random Forest Importance Table

Pre-COVID	*Importance measure*			
Variable	*Failed*	*Successful*	*MeanDecreaseAcc.*	*MeanDecreaseGini.*
Goal	24.25	86.94	87.49	759.76
Campaign length	32.92	48.44	61.55	604.64
Min	14.06	14.36	22.92	233.25
Q1	8.01	33.27	42.58	371.95
Median	4.51	28.23	34.88	309.80
Q3	9.40	29.32	38.33	365.11
Max	8.47	27.71	36.89	297.53
Number of tiers	54.77	74.36	99.73	759.87
Mean	5.40	30.22	37.68	420.13
Post-COVID	*Importance measure*			
Variable	*Failed*	*Successful*	*MeanDecreaseAcc*	*MeanDecreaseGini*
Goal	27.31	42.95	51.79	155.22
Campaign length	19.04	15.16	23.00	87.95
Min	0.20	10.43	10.21	41.23
Q1	-0.87	19.12	19.81	63.96
Median	-2.10	19.81	20.50	57.38
Q3	0.66	18.87	20.80	61.31
Max	-0.08	16.17	17.75	54.33
Number of tiers	29.07	38.43	47.52	132.75
Mean	-2.97	21.93	21.88	71.20

Appendix I - Data Source

https://webrobots.io/kickstarter-datasets/ : 2020-08-13 – CSV

Appendix J – Accuracy to K Size KNN Classification



Appendix K – Confusion Matrix KNN Classification

```
Confusion Matrix and Statistics

            Reference
Prediction   failed successful
  failed        287        171
  successful    213        329

               Accuracy : 0.616
                 95% CI : (0.5851, 0.6463)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.089e-13

                  Kappa : 0.232

 Mcnemar's Test P-Value : 0.03641

            Sensitivity : 0.5740
            Specificity : 0.6580
         Pos Pred Value : 0.6266
         Neg Pred Value : 0.6070
             Prevalence : 0.5000
         Detection Rate : 0.2870
   Detection Prevalence : 0.4580
      Balanced Accuracy : 0.6160

       'Positive' Class : failed
```