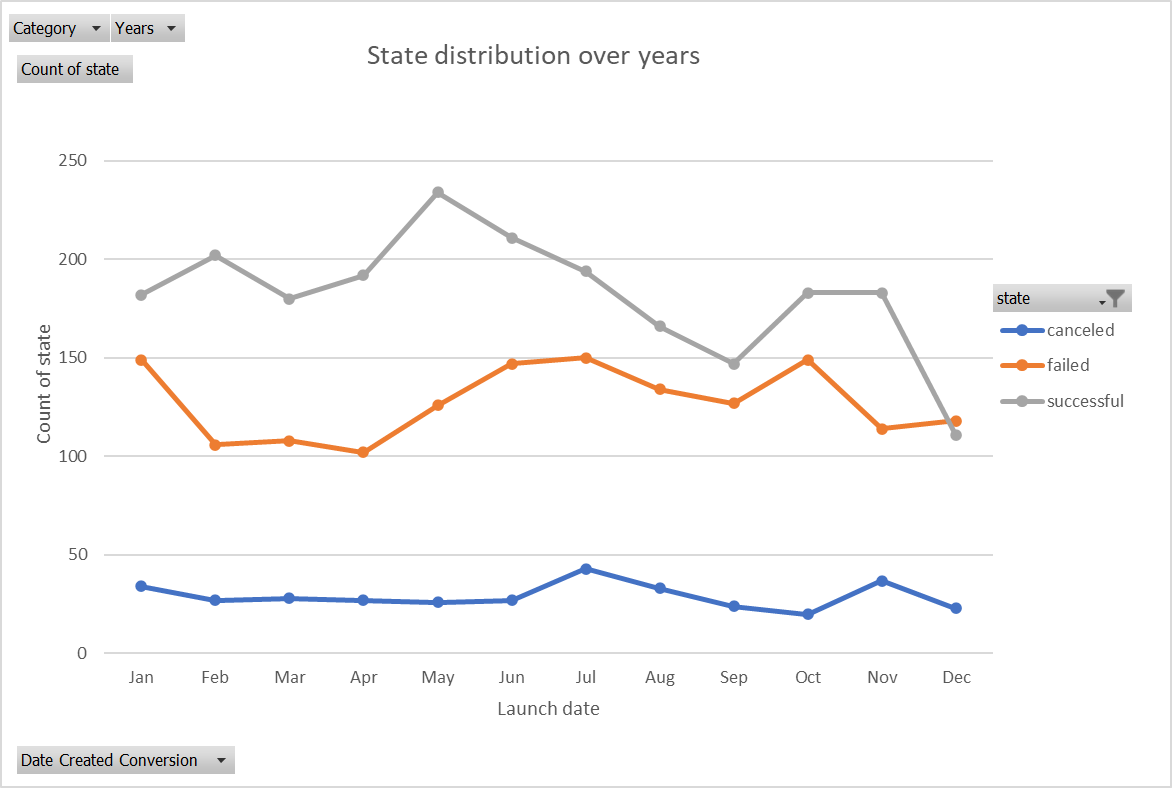
**Kickstarter Analysis Report**

**Dataset:**

The dataset used for the analysis consists of 4114 rows, which are projects proposed on the site. The columns that were on the dataset was the

1. Id
2. Name
3. Goal (the funding required)
4. Pledged (total funding received)
5. State (either successful, failed, live, or canceled)
6. Country
7. Currency
8. Deadline
9. Launch date
10. Staff pick (Kickstarter team-we love badge)
11. Backers count
12. Spotlight
13. Category /subcategory

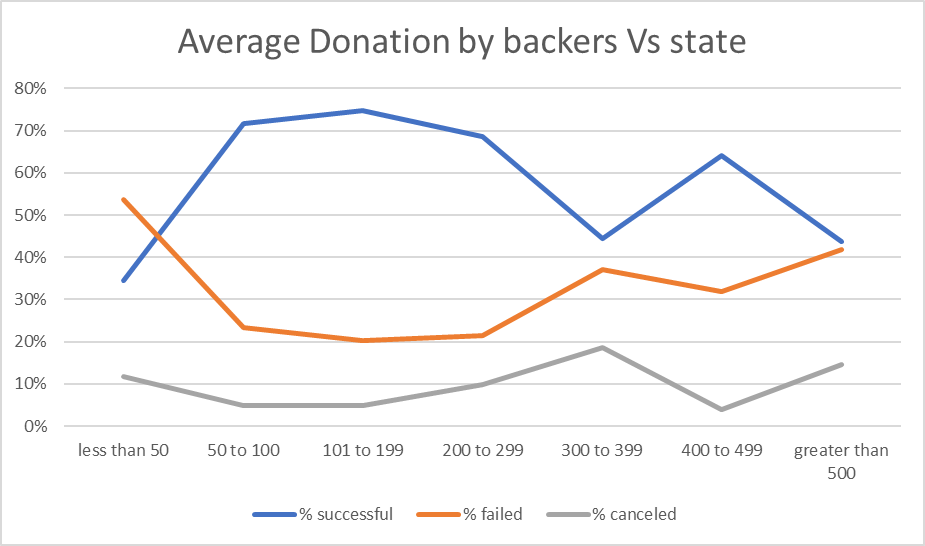
**Conclusions:**

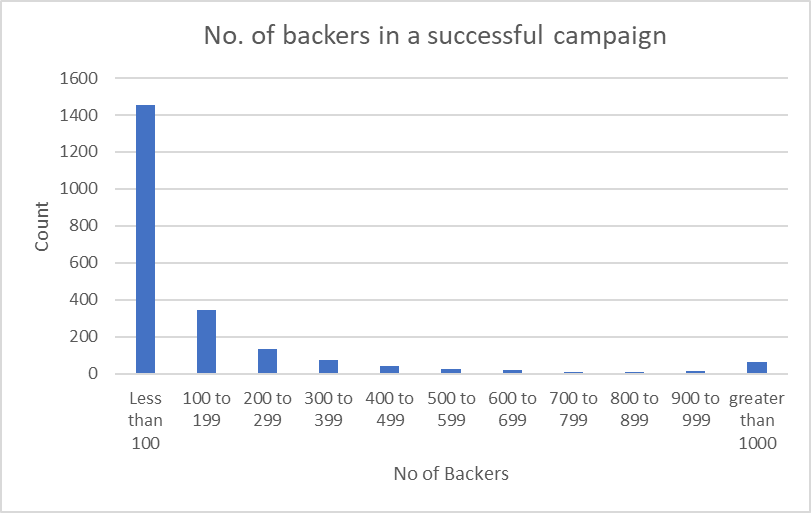
1. The success rates are higher for campaigns with lower goals. This can be visualized clearly from the line chart below. The failure and cancel rates increase with the goals increasing. From the analysis for the successful campaign, it is good to have goals within 5000, where the success rate is high. Between 5000 to 45000, the success and failure rates are almost similar, and above 45000, the failure and cancel rates steeply increase.
2. The projects launched in the first two quarters of the year are more successful than the last six months of the year. The higher number of successful goals are between May and June. This can be visually seen in the chart below.
3. Based on categories, music, theater, film, and video have high success rates. Interestingly, games, technology, and the food industry have significantly lower success rates on Kickstarter. Campaigns under specific subcategories are 100% successful, and the chart below presents them. Plays under the category theater have the highest number of total campaigns.

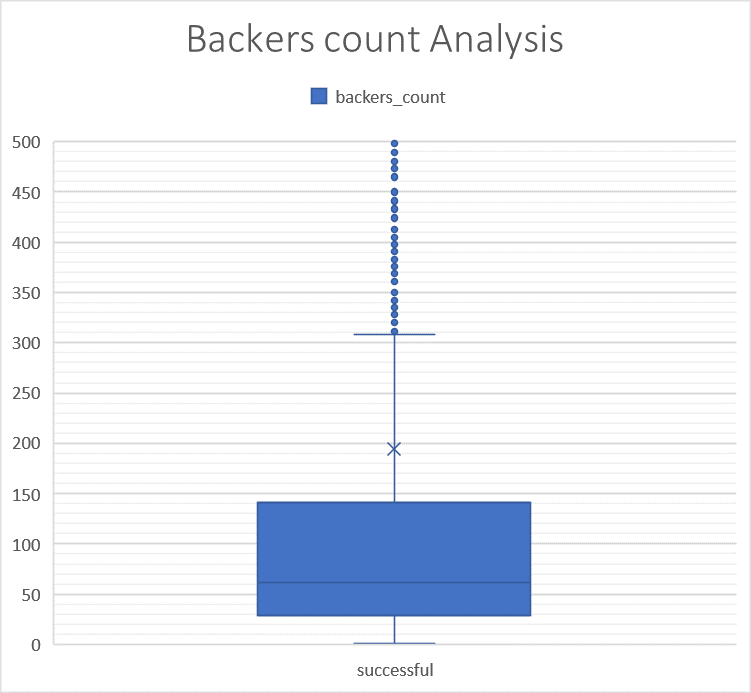
**Limitations:**

1. The dataset is minimal to provide insights on categories and subcategories that attracted a lot of backers. Each subcategory had very little data. When we filter the data based on countries, certain countries have very little data, limiting the analysis based on the geographical region.
2. The analysis is made on assuming that the goals are in USD, and when the goals are converted, then the goal-based outcomes may differ.

**Some other graphs and tables:**

1. Average donations by the backers is another useful that provides insights on the campaigns success. It is uncovered that the success rates are higher if the amount pledged by each backer lies between 50 and 300. Please find the chart that confirms the analysis.
2. On average, the number of backers must be around 195 for successful campaigns. The standard deviation is higher, and this might be the result of outliers in the dataset. A histogram can be plotted to analyze the spread of the data. It is found that it results in a skewed histogram with most of the data lying between 0 and 100. Box and Whiskers plot is also plotted to represent the outliers.



Many outliers in the dataset are the reason for increased standard deviation when analyzing the backers count.

**Prediction Model:**

After analyzing in excel, I understood the success rate depends on certain variables like the goal, staff pick, category, and subcategory. I used R to build a prediction model to help predict if the campaign will be successful based on the variables above.

The regular logistic regression model was built, and the model happens to be overfitting due to bias in the data and a lot of outliers.

As an alternative, I tried using decision trees to improve the model. The Classification and regression tree model had an accuracy of 85% on the test set. To improvise it further, I built a random forest model with a 1% increase in accuracy than the CART model.