An Analysis of Kickstarter Campaigns

Max Bowman & Bao Lam

1 Introduction

The goal of this analysis is to find trends in a relatively robust dataset of Kickstarter campaigns from 2009-2018 by analyzing charts and graphs to help better visualize the abundant amount of data. We settled on this dataset from the data science website Kaggle as it was well-maintained and had a copious amount of entries and distinct variables: In total, there were around 350,000 Kickstarter projects, grouped into exactly 159 different subcategories and 15 main categories. The dataset included variables such as the project's intended goal (`goal`); the amount raised (`pledged`); the state of the project (`state`), i.e., whether it was live, successful, failed, canceled, or suspended; the number of backers (`backers`); the project's main and sub-category (`main_category` and `category`) and other, less salient variables such as the goal and pledged amount adjusted for inflation. Due to its relative cleanliness and abundant amount of entries and variables, our dataset immediately prompted a variety of questions, the answers to some of which could possibly help people determine whether or not their project was a good fit for product – or if they were better off raising money through different means.

We found that one of the most interesting areas to draw conclusions from was the category that each project belonged to as well as its required funding. Using these variables we could make strong claims about, for example, which types of projects were more likely to become successful and which were merely a pipe dream. Furthermore, categories also helped explain some of the discrepancies we saw with outliers, like the ability for a project to go viral and get increased exposure. More refined research exposed trends such as which categories were more likely to have a large amount of backers but not necessarily produce the same amount of pledged dollars that another type of project may produce with less backers.

2 Data Cleaning

From the outset, our dataset seemed to be well-organized and maintained. There did not seem to be any glaring issues with the data (for example, a large percentage of missing values) and it seemed to be very intuitive despite a lack of a README.txt file (besides the brief description provided on Kaggle). Hence, the majority of our data-cleaning process consisted of removing unused variables, corrupted entries, and subsetting into new datasets for certain graphs so that the original would not be affected. In particular, we

- replaced any *NA* cells in the `name` column by changing them to ”Unknown”,
- removed corrupted or nonsensical entries, for example those with all three of the following conditions satisfied: 0 `backers` and an *NA* `amount_pledged` and `usd_pledged_real` ≥ 0,
- removed seven rows containing a questionable 1970 launch date (Kickstarter began in 2009) which had no statistically significant effect on the data,
- added columns for the amount of pledged dollars in thousands of USD (`pledged_thousands`) and the ratio between `pledged/goal`,
- removed variables that were of little use to us, such as `usd_pledged`, `usd_pledged_real`, and `usd_goal_real`,
- subsetted into new datasets for various charts; for example, we made subsets for the four largest categories in terms of pledged dollars, and the rest; a subset for strictly successful projects; and a subset for the six categories with the highest median `goal`, and
- for some of the data we stripped outliers from the dataset to show how it would affect our original analysis; however, due to such a large proportion of the data being considered outliers we only used that set for a minority of the graphs.

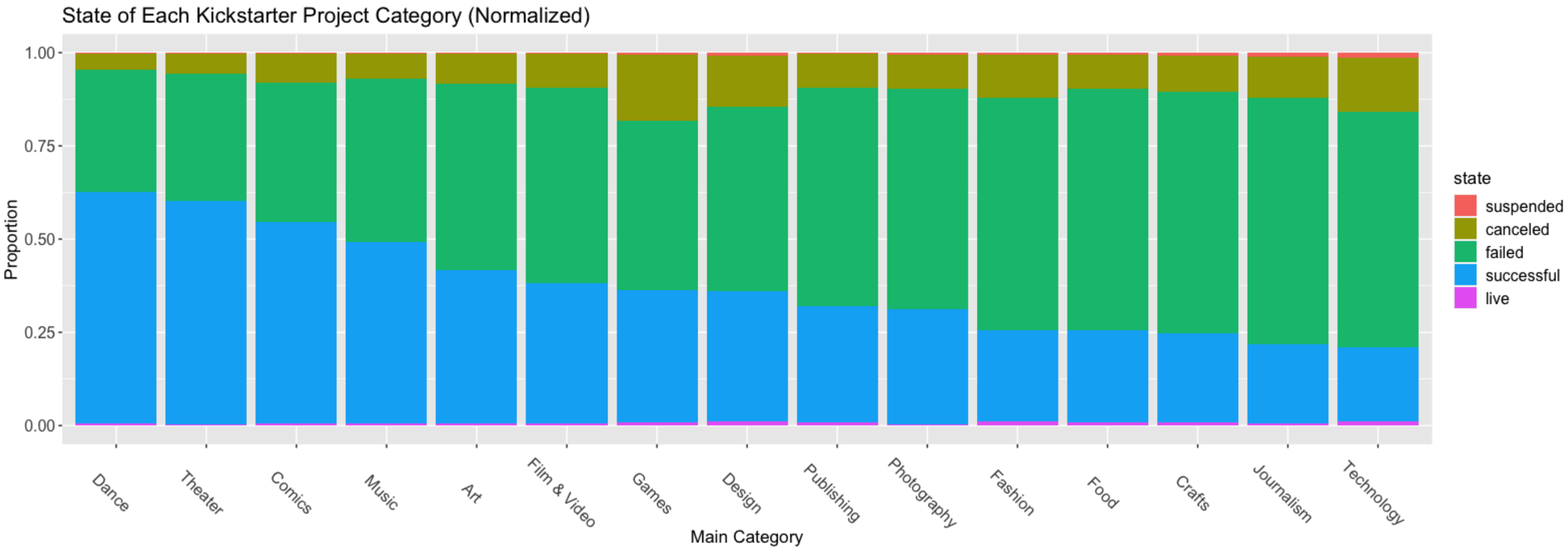

Figure 1: The State of Each Kickstarter Project Category (Normalized) [Final Iteration]

3 Analysis

3.1

Oftentimes when analyzing crowdsourcing campaigns, some of the immediate questions that arise are:

1. Which of the main categories has the highest success rate? Which has the highest failure rate? Why is this?

2. Is there a certain group of categories that tend to have the higher success rate? What about higher failure rate?

3. Do the more relatively popular categories tend to have higher failure rate or higher success rates? What about for more niche or unpopular categories?

4. Is there any relationship between `goal` and `state`?

5. Which type of products require the most amount of money to be successful?

6. Which type of products are most likely to fail after raising money?

7. Which country tended to have the most success?

From the barplot in Figure. 1, we seek to answer these questions. But first we comment on the graph itself.

Because there were an unequal amount of projects belonging to each category, we opted for a normalized, stacked bar plot rather than a side-by-side barplot (Figure. 2) because each category did not have the same amount of projects. Thus, in order to compare the success/failure rate, we would have to normalize the counts for each category. Also, for clarity we order the bars in descending success rate.

From the graph, observe that Dance projects had the highest success rate (around 60%) and lowest combined failure/cancel rate, while Technology projects had the lowest success rate (less than 25%) and highest combined failure/cancel rate. This immediately answers the first part of (1). Several questions then arise:

(i) Is the low success rate of Technology projects due to the fact that, from our intuition, technology projects often require significantly more funding compared to dance projects?

(ii) Because Technology projects tend to garner more attention and are more ”useful” compared to Dance projects, how could it be that Technology projects still have the highest failure rate?

In fact, the median `goal` for Technology projects is \$20,000 (thereby answering (5)) while for Dance projects it is a mere \$3,300. For part (i), one could argue that the amount of funding a project requires plays a significant role in either persuading or dissuading a potential backer: if, say, a Shakespeare- and audiophile-enthusiast was on the fence about whether they should fund a high-end Bluetooth headphone (\$72,000 goal) or a local production of Hamlet (\$8,000 goal) – both of which are newly-released – it seems they would much rather fund the latter project as their money is more likely to help that project come to fruition due to its lower cost. We will see later on in Section 3.2 that cost does indeed play a significant role in determining a campaign's success.

As for part (ii), perhaps the wider appeal of the Technology category is more a curse than a blessing. Because consumers often hope to fund the next iPhone or Fitbit, this creates an abundance of over-ambitious creators who seek to devise the next "It" gadget and net themselves a significant monetary sum in the process; But, whether due to lack of technical know-how or logistical issues, should they succeed in raising enough funds it is likely they will be unable to deliver on their product. Hence potential backers are dissuaded from contributing as they fear the creators overpromise and then under-delivery. This large amount of high-cost, partially-funded projects thereby saturates the market and greatly contributes to the failure rate of the Technology category.

Similarly, one can argue that categories such as Film & Video, Design (the two most popular Kickstarter campaigns, the Pebble Time smartwatch and the Coolest Cooler, belong to this category, although it could also be viewed as Technology) and Games are similar in that vein: they have widespread appeal but oftentimes require large monetary and technical funds. Oddly enough, however, from the graph we see that they have success rates in the middle of the pack, higher than, say, literary pursuits such as Publishing, Photography, and Journalism. It could be that these latter three categories, as they are more serious in nature than the entertainment categories, tend to garner less attention and attraction in a crowdsourcing website whose main focus is on tangible products that can be enjoyed by many.

On the other hand, from observation it appears that the niche categories oriented more directly toward the arts (Dance, Theater, Comics, Music, Art, Film & Video) tend to have higher success rates. As mentioned previously, perhaps one reason why is because they require lower funds and due to their being more popular than Publishing or Journalism are more likely to be funded. Hence our previous discussion answers (2) and (3).

For question (4), while we cannot directly answer this from the bar graph alone due to it not involving `goal`, from our intuition we can assume that Dance, Theater,

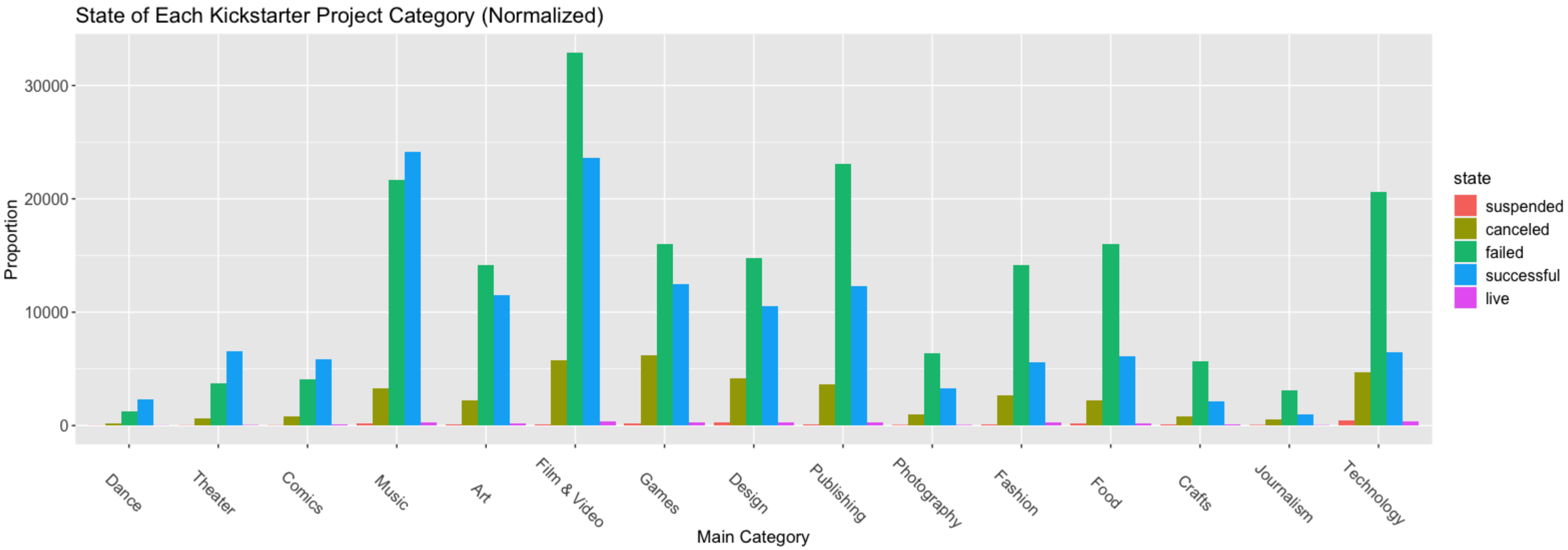

Figure 2: The State of Each Kickstarter Project Category (Normalized) [First Iteration]

Comics, and Music Kickstarter campaigns require the least funds. In fact, these categories are part of the five with the lowest median goal. One may then claim that there is a strictly negative correlation between goal and success rate, but as it turns out, this is not the case: Crafts projects have the lowest median goal yet maintain the third-highest failure rate. Perhaps this is because the four aforementioned categories tend to have a wider appeal, whereas Crafts campaigns, ranging from handmade Christmas ornaments to pens to knitted animals, are more niche (they account for around 2.3% of the total projects).

By looking at the relationship between goal, state, and main_category in the following section, we hope to further bolster our claims to questions (1)-(4).

3.2

The box plot in Figure 3. displays the distribution of the goals of the six Kickstarter categories with the highest median goals in descending order. The goal of this graph is to help bolster the previous claims made from our intuition by analyzing, for example,

- the spread of the goal per category,
- the median failure and success rate (category fixed), and
- the median goal of a successful (and failed) project per category.

Before the analysis, however, we comment on the graph itself. Note the interval of our y-axis; The range of the goal in the original dataset covers a wide interval, from $0.01 to $100,000,000. If we combine this with the observation that our graph is right-skewed (hence the reason why we compare median rather than mean, as skew significantly affects our mean), with the majority of the goal amounts clustered around $0 and essentially tapering off into the (hundreds of) thousands, then if we were to have graphed Figure 3. with the y-axis being [0, 100,000,000] instead of [0, 700,000], comparing the medians, IQR, and goal distribution of each category and state would have been extremely difficult, as shown in Figure. 4.

The following question then arises: Should we then graph goal with the outliers removed instead? The surprising detail to note from Figures. 3 and 4 is that because a significant amount of extreme goal values have many Kickstarter campaigns requiring that much amount of funding, there is a beautiful spread of (extreme outlier) values. The issue with creating box plots without the extreme outliers ($3 \cdot$ IQR) is that they are calculated to be those $>$ $60,000 (this accounts for 7% of our data), but in comparing Figure 5. (outliers removed) with 3 and 4, we see how much information is lost; while for Figure. 5 we can now better compare the medians and distribution, we cannot fully appreciate the entire spectrum of the ”extreme” values for each main_category and whether or not the projects associated with those goal values succeeded or failed. Hence we claim that our graph in Figure 3. is a good compromise, as it preserves the pattern of the extreme entries with goals $\geq$ \$700,000 while also not significantly altering the median values in the original dataset and by allowing us to better compare the medians, IQR, and range (i.e., the boxes are not completely squished).

As for the analysis, from the graph we can see that the most expensive categories tend to be the entertainment-oriented, in other words, those with some of the lowest success rates. This supports our claim in Section 3.1 that the more costly main_category tended to have lower success rates. In fact, for the first four categories there were numerous campaigns across the entire spectrum that either failed or were canceled. On the other hand, the latter two categories, journalism and publishing (both with a \$5,000 median goal), had some of the lowest success rates from 3.1, although admittedly their goal values were not spread out as much. Hence it could be that the high goal combined with their relative lack of entertainment and appeal contributed to their high failure/canceled rate; for these two categories, only for low-budget goals were there many successful projects, but even in this monetary spectrum there were still many failures.

To help answer (4) from Section 3.1, note from the significant spread for each state that the higher the goal the greater the possibility that the project will fail or be canceled – while the success rate slowly tapers off – regardless of category. Hence we claim that the higher the goal, the more likely the project is to fail or be canceled. Again, this could be because the project is high-budget, it is unlikely that the funds can be raised in the allotted time due to the technical and logistical issues discussed in Section 3.1 as well as the fact that because the first four categories are so popular there may be many projects that a potential donor may to decide to back, and thus there is a greater amount of competition between campaigns. A big-budget project is hence less likely to be successful if they lack dedicated creators who possess both the technical expertise and marketing savviness that will allow them to not only appeal to potential backers but also deliver on their promises.

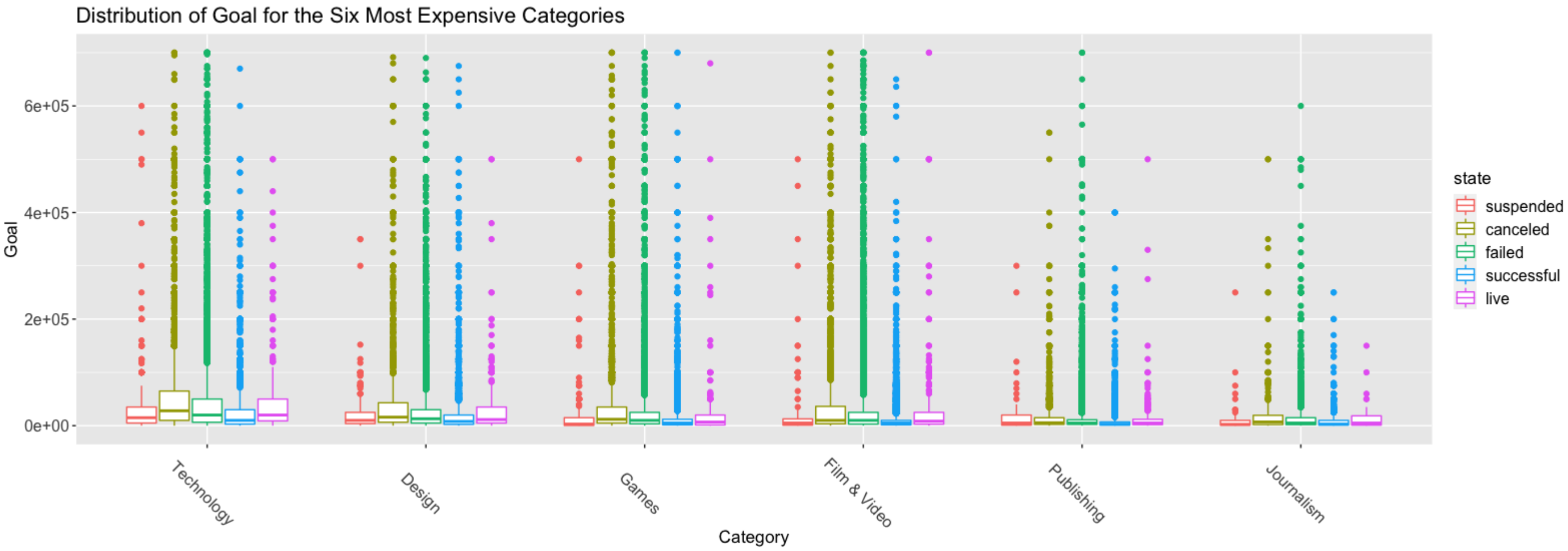

Figure 3: Distribution of the Goal of the Six Most Expensive Categories [Final Iteration]

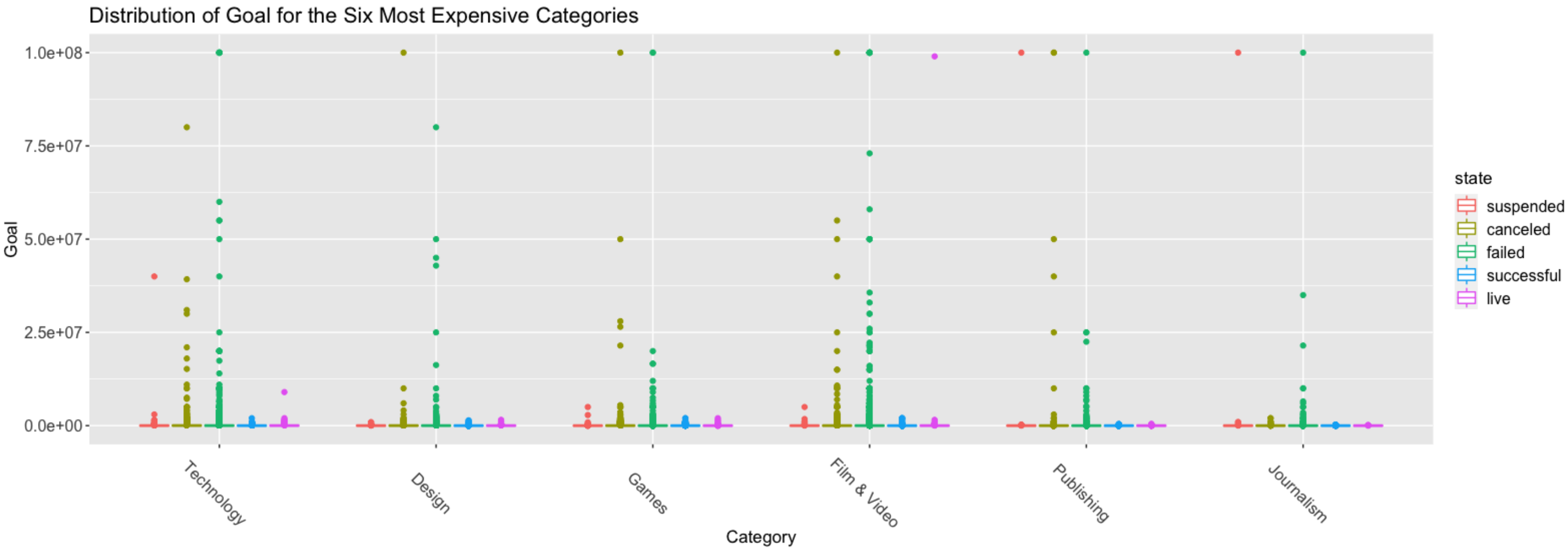

Figure 4: Distribution of the Goal of the Six Most Expensive Categories [First Iteration]

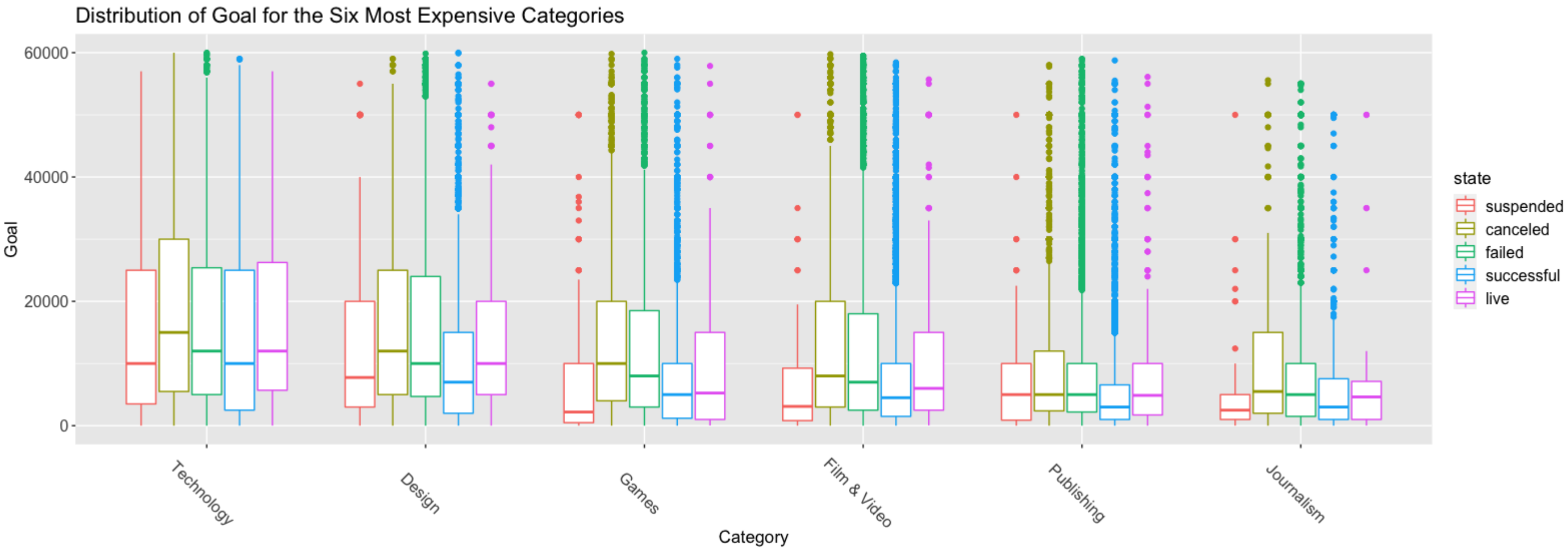

Figure 5: Distribution of the Goal of the Six Most Expensive Categories [Second Iteration]

3.3

Another important segment of the data was the amount of pledged dollars (pledged_thousands) to the project and the amount of people backing the project (backers). There was only a little bit to do with the pledged numbers needing to be reduced to thousands so that it was easier to read, and the removal of the variable usd_pledged (by making a subset of all other variables), so there was no confusion on which we were using going forward. Before the plot was made we knew the amount of backers versus the amount pledged would bring some interesting findings to light but we really wanted more of a definitive answer as to what that relation was. When the plot was first created, all the categories were located on one plot but the scale of things were very skewed. The trend lines in Figure. 7 reached pledged numbers of at least 4.5 million (4500 thousand as shown) and reached up to 12 million dollars pledged. These, compared to the paltry figures in the rest of the pack, were easily recognized as different and were thus separated from the group to form their own graph. Please note that the two graphs have different y-axis scales in an attempt to make the graphs more easily readable. In addition, when delving deeper into this finding, it becomes clear that the top four have one main common trait: mass appeal. All four of these could reach almost every demographic, and by default that makes these categories have the largest possible donor constituency. Things like art, and theater will by default only appeal to those more niche donors and are also much more likely to be locally based. This coupled with a larger starting goal on average (Figure 8) for the largest four, make it very clear just how much more lucrative those categories are than the rest.

Also interesting is the shape of each of the curves. Some appear linear, while others are more logarithmic. Even between the categories "games" and "design", the differences are evident. For design, its curve indicates a higher spending backer group with, on average more pledged dollars per backer than games. Games on the other hand has more widespread support with a larger, faster increasing backer pool, but less pledged per backer. This can speak to the type of people backing each of the types of category. Another conclusion that can be drawn from those discrepancies is the fact that some products may cost more than others to fund. For example, something in the design category is most likely going to cost much more than a song in the music category. This will cause less backers to be needed for a larger amount of money to be raised. As shown in Figure. 6, music is relatively highly backed but the amount pledged is lower due to the smaller cost associated with producing that good as opposed to a high cost, high pledge category like "Film & Video" (Figure. 7). The rationale applied to these categories can be applied to any of the categories from "Food" all the way down to the smaller ones like "Theater". Overall to answer the

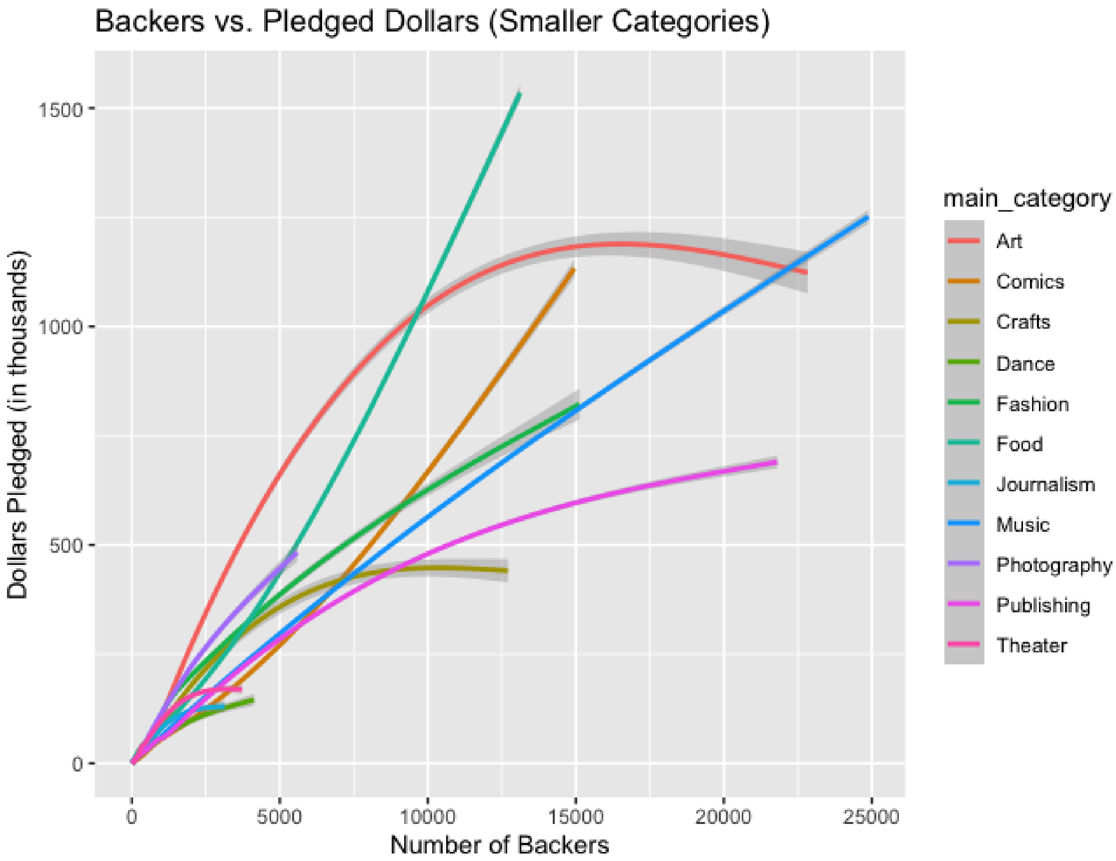

Figure 6: Backers per Donation Smooth (Largest 4)

question we started with, there is definitely a relation between the variable `backers` and `pledged_thousands` but the stronger determining force is again the category in both shape and amount of both variables.

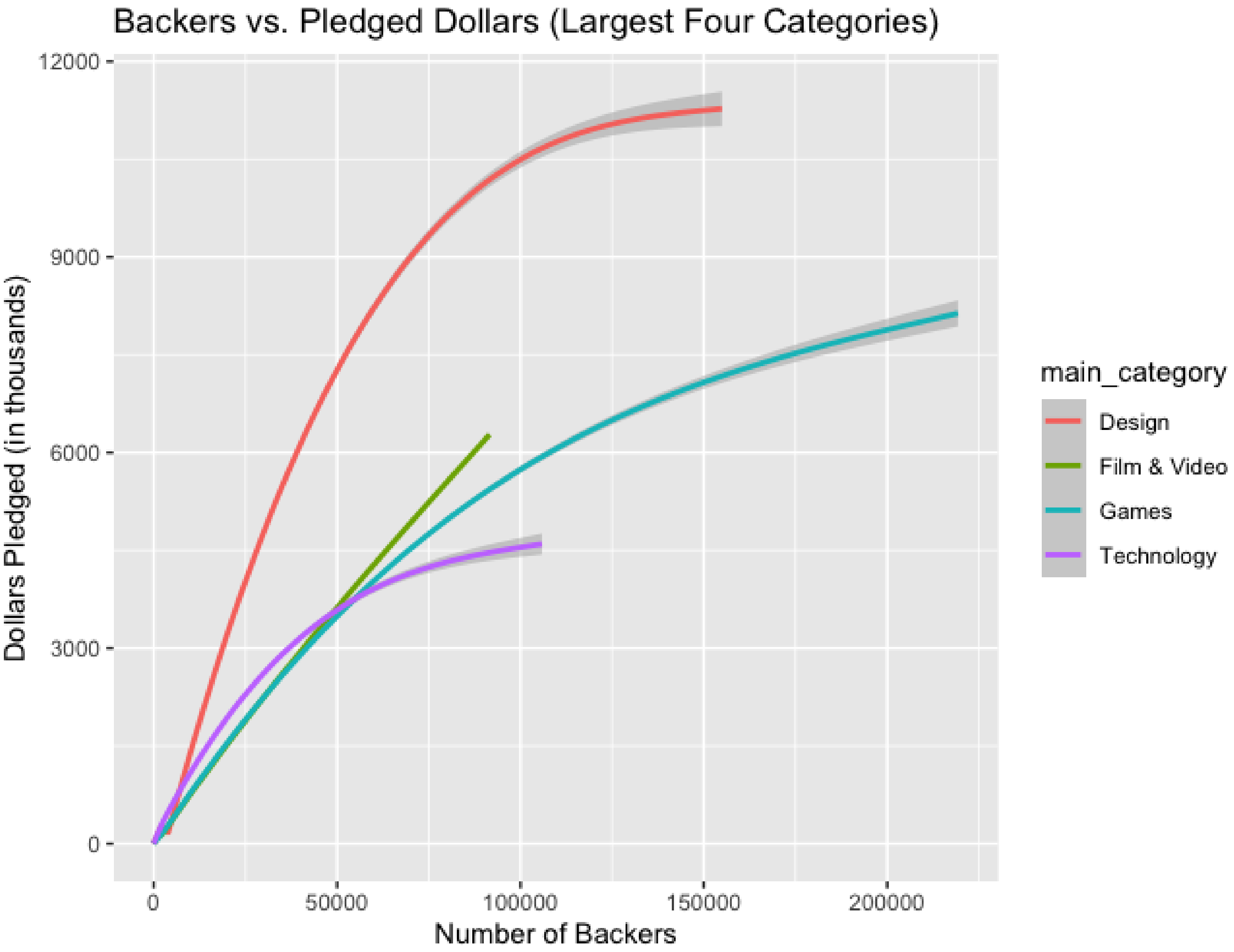

Figure 7: Backers per Donation Smooth (Remainder)

3.4

The dataset also came with timestamps of when the projects were started down to the second. This area required some cleaning with exactly 7 rows having a start time during the year of 1970. Due to some odd results we decided to test this plot when outliers were removed. For other plots, there was not really a reason to, but with time being a variable in this plot it was best that we did it to get a more central feel for the data. We did not want to overcompensate though and as such we only removed extreme outliers by making a subset without them. This gave us more accurate information, and kept the plot closer to what the averages would be. With these adjustments, we were able to find an interesting set of observations that show the trend in goals over time. The largest increase was for technology, which went from a range of around 3000 to over 20000. This means by 2018 our best guess at a goal for a technology project would be 20000 dollars. This was an outlier of a plot all by itself with only games coming close to the rapid increase in necessary funding. We attributed this to a couple of factors including the larger overhead for technology based projects as opposed to some of the other, cheaper to produce goods. This was sort of a surprise as we thought that by removing those outliers from the data the curve would be a little less drastic in change over time. It represents a more middling part of the data making it more accurate for the smaller scale projects and thus gives us the insight that not only were the large scale kickstarters demanding a high goal and price to pay for its pool of donors, but almost all tech, design and even games were increasing their respective floors for the amount produced. Surprisingly enough, most of the categories were stagnant in their goals with ones like the ”comics” category even approaching a decline.

These trends show how big of an undertaking the starting of a project in each category has had over time. Music more or less costs the same to produce, accounting for inflation, as it did in 2010 and thus the goals for the projects only increased slightly. Technology on the other hand has included larger and larger undertakings, with it having a massive increase in goals as time went along. As mentioned previously in section 3.1 with Figure. 1, categories like Theater and Dance had the highest rate of success. Using that knowledge combined with the findings of this graph explained what was at first a confusing conclusion from Figure. 1. Categories play a huge role in determining goal and pledged amount and the category also is a big determining factor in how static the goals stay. Lower goals may mean higher success rate, but that includes extremely small goals from small scale projects that only really appear (in any significant amount) in the smaller scale categories mentioned in 3.3. The real looming question that we hoped to answer from this analysis was did goals of projects as a whole increase over time, as expected. The short answer is no. With

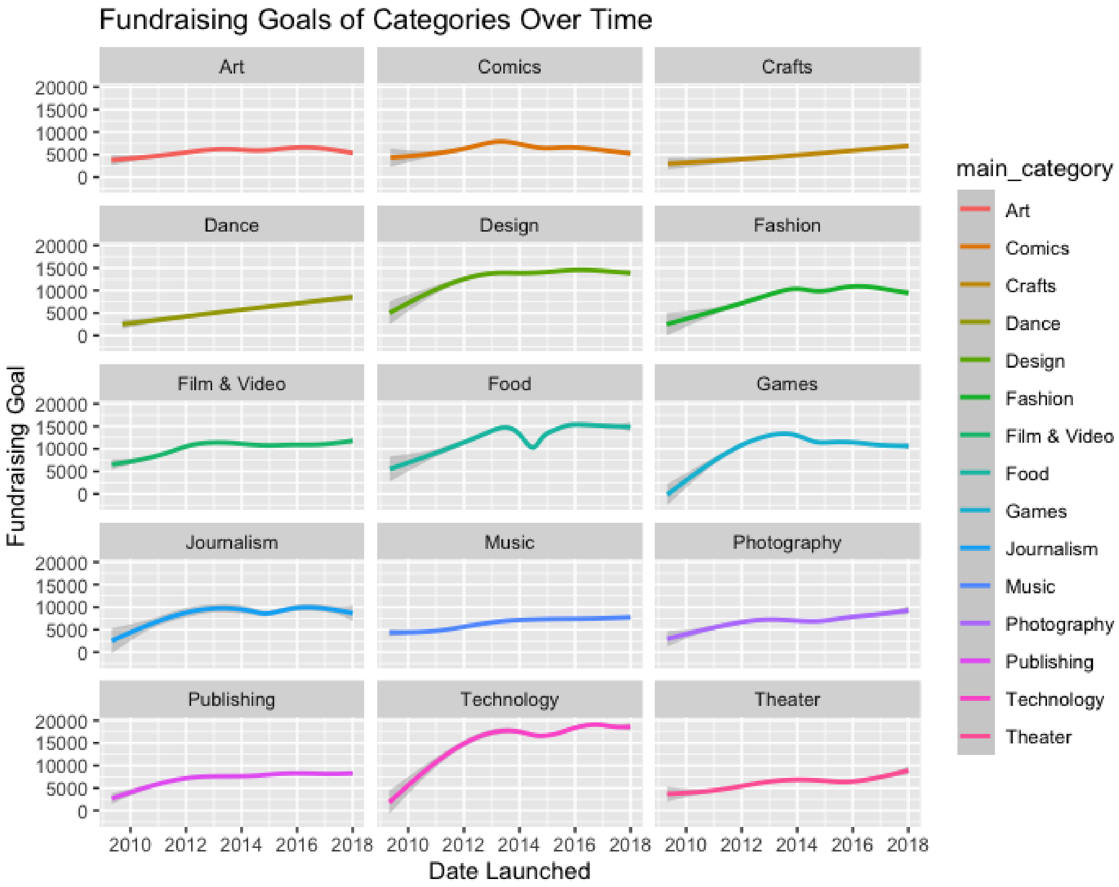

Figure 8: Trend line for fundraising goals as a function of time for each main category

slight increases in some and massive jumps in others, from 2010 to 2018 there really seems to be only a very slight correlation between the goals of projects and time that the project launched. Technology being the industry it is, we may expect it from that category but otherwise as a whole, projects did not increase their goals with the times.

4 Conclusion

This dataset presented challenges that were not present in an initial analysis of the raw data but were exposed through trial and error when it came to creating a sound report for the data. Questions that we did not anticipate grew out of new variables that we created and graphs that did not end up looking exactly as we anticipated or hoped. The cleaning process was all in all not a very grueling one with this dataset being very well maintained and curated from the start. There were however roadblocks when it came to analysis and the vast amount of zeros resulting from countless failed projects. We overcame this by using proportional measures and also trend lines that gave good approximations and made the graphs readable. One finding is for sure: the category and required funding of the project is a heavy determining factor in what kind of data it produces. For example, we did not expect a category as relatively obscure as Dance to have such a high margin of success; on the other hand, however, we did expect for Technology projects to have the lowest success rate, due to its high cost and abundance of competition.

For determining the relationship between backers and the pledged amount, both logarithmic and linear trends were shown with their own unique connotations. We cannot know for sure, but the price of the products and scalability are both factors in the size and shape of the plots. As for goals over time, the trends are not purely categorical as there was slight increase over all categories, but again project overhead and increased notoriety are good ways to summarize how these trends were formed.

There were questions that unfortunately could not be answered with our analysis. For example: Which type of products are most likely to fail after raising money? Because the required funding varied dramatically for each category, it was difficult to set a threshold for how much could be raised before we could take the product into consideration. Another question that was left unanswered was ”Which country tended to have the most success?” Initially, we had believed that the currency was associated with the creator’s country of origin, but as it turns out from later research this was not always the case. For example, creators in Australia could opt for raising funds in USD rather than AUD.

Nevertheless, from this plentiful and well-maintained dataset we were able to come up with thoughtful analysis that could not immediately be inferred from first glance. In the future, perhaps we could gather more variables in order to make deeper connections and claims. For example, we could request the creator’s country of origin; the quickness in which funds are raised before the deadline (to gauge popularity); or a rating scale from potential backers as to how favorable they view the creator’s campaign (to signal to donors whether or not others believe this project is legit and will follow up on their promises). Crowdsourcing is the future, and with its maturing hopefully new and refined datasets will be released so that further analysis can be conducted.

5 References

- Mouillea, M. (2018, February 8). Kickstarter Projects. Retrieved April 6, 2020, from *https://www.kaggle.com/kemical/kickstarter-projects*

- Prabhakaran, S. (2016, December 9). Outlier detection and treatment with R. Retrieved March 29, 2020, from *https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/*

6 Code Appendix

6.1 Initialization

```
library("ggplot2")
ks <- read.csv("Kickstarter2018.csv")
```

6.2 Cleaning

```
# Checking which specific rows have NA values and total NA values
sum(is.na(ks)) # 3801 NA entries
sum(is.na(ks$ID))
sum(is.na(ks$name)) # 4 NA entries
sum(is.na(ks$category))
sum(is.na(ks$main_category))
sum(is.na(ks$currency))
sum(is.na(ks$deadline))
sum(is.na(ks$goal))
sum(is.na(ks$launched))
sum(is.na(ks$pledged))
sum(is.na(ks$state))
sum(is.na(ks$backers))
sum(is.na(ks$country))
sum(is.na(ks$‘usd pledged‘)) # 3797 NA entries
sum(is.na(ks$usd_pledged_real))
sum(is.na(ks$usd_goal_real))

# Remove specific corrupted/nonsensical entries
# For example, NA pledged amount AND 0 backers AND pledged amount in USD >= 0
```

```
ks <-ks[!(ks$backers == 0 & is.na(ks$usd.pledged) & ks$usd_pledged_real >= 0), ]

# Remove variables/columns that we will not use
ks <- subset(ks, select = c(ID, name, category, main_category,
                            currency, deadline, goal,
                            launched, pledged,state, backers,
                            country, ratio, pledged_thousands))

# Eliminate the entries with 1970 launch dates
ks <- ks[!ks$ID %in%
           c(1014746686, 1245461087, 1384087152, 1480763647,
             330942060, 462917959, 69489148), ]

# Assign a variable to show which rows contain null values
# in the specified column
# Create a variable to print the rows from the name column
DF <- read.csv("2018.csv", na.strings=c("NA", "NULL"))
new_DF <- subset(ks, is.na(ks$name))

# Changing the NA rows found using previous code
ks$name[is.na(ks$name)] <- "Unknown"

# Create a new variable for ratio of pledged dollar amount
# over project goal
ks$ratio <- (ks$pledged/ks$goal)

# Create an additional column to make the graph more compact
# for pledged dollars in thousands of dollars
ks$pledged_thousands <- (ks$pledged/1000)
```

6.3 Main Code

```
# Factor so as to order the bar plot
# in decreasing success rate
ks$state <- factor(ks$state, levels = c("suspended", "canceled",
                                        "failed", "successful", "live"))
ks$main_category <- factor(ks$main_category,
                           levels = c("Dance", "Theater",
                                           "Comics", "Music",
                                           "Art", "Film & Video",
                                           "Games", "Design",
                                           "Publishing", "Photography",
                                           "Fashion", "Food",
                                           "Crafts", "Journalism",
                                           "Technology"))

# Graph 1 (Bar plot): Help visualize success/failure rate per category
# Initial version
ggplot(data = ks) +
  geom_bar(mapping = aes(x = main_category, fill = state), position = "dodge") +
  xlab("Main Category") +
  ylab("Proportion") +
  ggtitle("State of Each Kickstarter Project Category (Normalized)") +
  theme(axis.text.x = element_text(angle = -45),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 13),
        plot.title = element_text(size = 16),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 12))

# Graph 1 (Bar plot): Help visualize success/failure rate per category
# Final version
ggplot(data = ks) +
  xlab("Main Category") +
  ylab("Proportion") +
  ggtitle("State of Each Kickstarter Project Category (Normalized)") +
  geom_bar(mapping = aes(x = main_category, fill = state), position = "fill") +
  theme(axis.text.x = element_text(angle = -45),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 13),
        plot.title = element_text(size = 16),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 12))

# Indices of ks of the six categories
# with the most expensive goals by median
```

```
most_expensive_category <- which((ks$main_category == "Technology" |
                                  ks$main_category == "Publishing" |
                                  ks$main_category == "Journalism" |
                                  ks$main_category == "Games" |
                                  ks$main_category == "Film & Video" |
                                  ks$main_category == "Design"))

# Subset those entries
most_expensive <- ks[most_expensive_category, ]

# Factor these categories so that
# box plot is listed in decreasing median goal values
most_expensive$main_category <- factor(most_expensive$main_category,
                                       levels = c("Technology", "Design",
                                                  "Games",
                                                  "Film & Video",
                                                  "Publishing",
                                                  "Journalism"))



# Get the quantile and IQR of goal
Q <- quantile(ks$goal, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(ks$goal)

# Entries of ks without extreme outliers
no_outliers <- subset(most_expensive, (most_expensive$goal > (Q[1] - 3 * iqr)) &
                        (most_expensive$goal < (Q[2] + 3 * iqr)))

# Graph 2 (Box plot): Helps analyze relationship
# between goal, state, and main_category
# No outliers
ggplot(data = no_outliers,
       mapping = aes(x = main_category,
                     y = goal, color = state)) +
  geom_boxplot ()+
  xlab("Category") +
  ylab("Goal") +
  ggtitle("Distribution of Goal for the Six Most Expensive Categories") +
  ylim(0, 60000) +
  theme(axis.text.x = element_text(angle = -45),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 13),
        plot.title = element_text(size = 16),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 12))

# Graph 2 (Box plot): Helps analyze relationship
# between goal, state, and main_category
# Final iteration
ggplot(data = most_expensive,
       mapping = aes(x = main_category,
                     y = goal, color = state)) +
  geom_boxplot ()+
  xlab("Category") +
  ylab("Goal") +
  ggtitle("Distribution of Goal for the Six Most
          Expensive Categories") +
  theme(axis.text.x = element_text(angle = -45),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 13),
        plot.title = element_text(size = 16),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 12))


# Graph 2 (Box plot): Helps analyze relationship
# between goal, state, and main_category
# Final iteration (range changed)
ggplot(data = most_expensive,
       mapping = aes(x = main_category,
                     y = goal, color = state)) +
  geom_boxplot ()+
  xlab("Category") +
  ylab("Goal") +
  ggtitle("Distribution of Goal for the Six Most Expensive Categories") +
  ylim(0, 800000) +
  theme(axis.text.x = element_text(angle = -45),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 13),
        plot.title = element_text(size = 16),
        legend.title = element_text(size = 13),
        legend.text = element_text(size = 12))


# Barplot of only the successful projects for each category
# Helpful in visualizing data
successful_subset <- subset(ks, state == "successful",
                            select = (c(ID, name, category,
                                       main_category, goal, pledged,
                                       state, backers)))
ggplot(successful_subset, aes(x = factor(main_category),
                              fill = main_category))+
  geom_bar() +
  labs(x = "Main Category", y = "Number of Successful Projects",
       title = "Number of Successful Projects Per Main Category")

# Plot covering all main categories in relation to
# pledged dollars and backers
ggplot(ks, aes(x = backers,
               y = pledged_thousands,
               color = main_category)) +
  geom_smooth()

# Making subsets of all the 4 largest categories based on a test plot
# A plot covering only the top 4 categories in terms
# of pledged dollars and backers
big_main_category <- subset(ks, main_category == "Technology" |
                              main_category == "Design" |
                              main_category == "Games" |
                              main_category == "Film & Video",
                            select = (c(ID, name, category,
                                        main_category, goal,
                                        pledged, state, backers,
                                        pledged_thousands)))
```

```
ggplot(big_main_category, aes(x = backers, y = pledged_thousands,
                              color = main_category)) + geom_smooth() +
  labs(x = "Number of Backers", y = "Dollars Pledged (in thousands)",
       title = "Backers vs. Pledged Dollars (Largest Four Categories)")


# Create subset of all smaller categories to compliment the other plot
# Plot covering the rest of the main categories
small_main_category <- subset(ks, main_category != "Technology" &
                                main_category != "Design" &
                                main_category != "Games" &
                                main_category != "Film & Video",
                              select = (c(ID, name, category,
                                          main_category, goal, pledged,
                                          state, backers, pledged_thousands)))

ggplot(small_main_category, aes(x = backers,
                                y = pledged_thousands,
                                color = main_category)) +
  geom_smooth() +
  labs(x = "Number of Backers",
       y = "Dollars Pledged (in thousands)",
       title = "Backers vs. Pledged Dollars (Smaller Categories)")

# Obtaining the quartiles and IQR data
Q <- quantile(ks$goal, probs = c(.25, .75), na.rm = FALSE)
iqr_ks <- IQR(ks$goal)

# Removing outliers from the dataset
no_outliers <- subset(ks, ks$goal > (Q[1] - 3*iqr_ks) &
                        ks$goal < (Q[2] + 3*iqr_ks))

# Creating a dataset of only Technology projects without outliers
# Plot for tech only set for testing non outlier application
no_outliers_tech <- subset(no_outliers, main_category == "Technology")
ggplot(no_outliers_tech, aes(x = launched,
                             y = goal,
                             color = main_category)) +
  geom_smooth() +
  facet_wrap(main_category ~ ., ncol = 3) +
  labs(x = "Date Launched", y = "Fundraising Goal",
       title = "Fundraising Goals of Categories Over Time")


# Plot with outliers included
# Used for comparison between using outliers and not using outliers
# for this specific analysis
ggplot(ks, aes(x = launched, y = goal, color = main_category))
+ geom_smooth() + facet_wrap(main_category ~ ., ncol = 3)
+ labs(x = "Date Launched", y = "Fundraising Goal",
       title = "Fundraising Goals of Categories Over Time")

# Plot without outliers
ggplot(no_outliers, aes(x = launched, y = goal, color = main_category))
+ geom_smooth() + facet_wrap(main_category ~ ., ncol = 3)
+ labs(x = "Date Launched", y = "Fundraising Goal",
       title = "Fundraising Goals of Categories Over Time")
```