Factors Associated with Kickstarter Success

Introduction

Kickstarter is a global crowdfunding platform focused on creativity and it could be a good way to generate start-up funds and advertise your business prior to opening. While many projects are funded successfully, more projects fail generally. It's a smart choice for creators to understand factors associated with Kickstarter success before investing time in a Kickstarter campaign.

This report explores what generally makes for a successful campaign using past Kickstarter campaign data and then predicts the chance of success for a new community-based business aiming to raise at least $25,000 from at least 1000 backers in Detroit. Our analyses conclude that the funding goal, the length of project name, the month and weekday of the deadline, the year, month and weekday of creating date, and the time span from launching the campaign to deadline have impacts on the Kickstarter success. The chance of success of the new community-based business is predicted as 0.665.

Methods

There are two main goals for our analyses. First, to determine generally what makes for a successful campaign using past Kickstarter campaign data; Second, to help our client to predict the chance of success for their new community-based business aiming to raise at least $25,000 from at least 1000 backers.

To address these two goals, a Logistic Regression model was constructed to explore the association between the Kickstarter success and other factors and then predict the probability of success for the client's project. *SuccessfulBool*, a bool variable provided in the data to indicate whether a campaign is successful or not, was used as the response. Based on the corresponding relationship between state and *successfulBool* given in the data, all campaigns with live, suspended, canceled, and failed states were treated as NOT successful in *successfulBool* variable. Since what we care about more is how to be successful, we used *successfulBool* instead of state in our analyses. The problem was a binary classification, and we aim to explore the how other factors impact the classification and predict the probability of success for a given campaign. The data and the goal perfectly aligned with Logistic Regression model.

Since the client's campaign will in USD and the client is more interested in the analysis on US campaigns, our analyses mainly focused on the project data from US instead of all data. The data set was large enough and we had plenty of US campaign data, so excluding data from other countries to do a sub-analysis would cause little information loss in terms of our goals and could make our analyses more targeted.

Our Logistic Regression model took the *successfulBool* as response, and funding goal, length of project name, length of blurb, whether communication is disabled by creators, month, day and weekday of deadline, year, month, day, and weekday of project created at, number of days from creating to launching, number of days from launching to deadline and category of project as predictors.

Here are details on reasons why we excluded or split some variables.

In our model, we excluded the amount of money a project has raised (*pledged*) and amount of people having supported the project (*backers*). Whether a campaign is successful will influence the money a project has raised, and the number of people having supported to the project. It would be a circular logic if we include these two variables. Furthermore, there is no way for creators of projects to control the amount of money having been raised and the amount of people having supported, so including these two variables makes no sense for our goal and model. We also excluded the name and blurb of a project due to skill limit on text analysis, and we only used the length of project name and the length of blurb to provide some information on name and blurb.

For the time variables including deadline given for successful funding (*deadline*), state changed when campaign went to success or failure (*state_changed_at*), time the project was created at (*created_at*), time the project was launched at (*lauched_at*), we excluded the time *state_changed_at*, which was almost perfectly correlated with *deadline*. For the other three, to minimize the association among factors, we only included the year, month, day, and weekday of *created_at* and the month, day, and weekday of *deadline* based on the correlation calculation. To depict the timespan, we also included the time from creating to launching (*create2launch*) and the time from launching to deadline (*launch2deadline*).

The significance level of predictors in the Logistic Regression model was used to determine what generally makes a successful campaign: significant predictors were factors having impacts on the success and the regression coefficients of these significant predictors were used to depict how they impacted the success.

One complexity was about the imputation of unknown values of predictors in our model for the client's project. For the client's project, we could only set the funding goal as $25,000 and the year the project will be created at as 2022. To predict the probability of success, we used the best possible values from previous projects to impute the values of other predictors. For a non-significant predictor, we just used the average of that predictor or the mode of that predictor if it is categorical in previous projects. For a significant predictor, if the regression coefficient was negative (the predictor had negative effect on probability of success), we used max value of that variable in previous projects; if the coefficient was positive (the predictor had positive effect on probability of success), we used min value of that variable in previous projects.

There were two main limitations in our analyses.

One was that we could not constrain the number of backers to be at least 1000. We only had information on the number of backers having support, which was more like an outcome variable rather a controllable factor. And as explained above, due to circular logic, we did not use that variable in our Logistic Regression model. So, we only made prediction on the probability of success for a project created in 2022 and aiming to raise $25,000. And to satisfy the client's specific requirement on number of backers, we split the variable *backers* into two groups (value of 1 *backers* is larger than or equal to 1000; value of 0 if *backers* is less than 1000) and constructed another simple Logistic Regression model for projects successfully raising at least $25,000 using all variables we used in the Logistic Regression model for explaining and predicting probability of campaign success to predict the probability of having over 1000 backers for a project that successfully raised over $25,000.

The other limitation was that the data set was imbalanced. Only 31% of projects were successful, and only 6% of projects successfully raised at least $25,000. The imbalanced data made the model tend to classify a project as not successful and then the probability of a project successfully raising at least $25,000 would be under-estimated. Since the quantity of data was sufficient, we down sampled the not-successful projects to be the same number as the successful project to construct a new balanced data set.

Results

Each row of the data represents one campaign/project on Kickstarter. The initial data set contained 20,632 observations and 68 variables. We took 25 variables (as shown in Table 1) related to the client's concern into account and 14,138 observations (omitted 3 observations with NA values) with country value of US and currency value of USD were used to help perform sub-analysis on US campaigns as the client requested. Upon investigation, 540 observations with extremely high funding goal, amount of money having raised, number of backers or number days from creating to launch were removed as outliers, and 13,598 observations were remained. Table 1 shows the summary on statistics for 25 variables (some details are omitted due to space limit.)

Variable	Definition	Mean	Standard Deviation	Median (IQR)
project	Project ID	13598 unique values (1 value for 1 observation)		
name	Project Name	13590 unique values; 8 names are duplicated twice.		
state	State of a project	successful: 4168; canceled: 1591; failed: 7408; live: 295; suspended: 136.		
*success**	A project is successful or not	successful (1): 4168; not successful (0): 9430.		
*category**	Field a project is on	25 categories (Details in appendix).		
*funding_goal**	Funding Goal (USD)	35911	62218	13056 (4500, 42000)
*name_len**	Length of Project Name	5.98	2.79	6 (4, 8)
*blurb_len**	Length of Blurb	13.01	3.21	15 (8, 23)
pledged	Amount of money having been raised	15014	44092	911 (37, 492204)
backers	Number of people having support the project	142.50	479.86	14 (2, 72)
*create2launch**	Number of days from creating to launching for a project	41.69	70.55	14 (4, 45)
*launch2deadline**	Number of days from launching to deadline for a project	34.66	11.90	30 (30, 40)
*disable_communication**	Creator of a project disabled the communication or not	True: 136; False: 13462.		

Time Variable	Definition	Mean	Standard Deviation	Median (IQR)	Mode

deadline_year	Year of deadline (2009 - 2017)	2015	1.37	2015 (2014, 2016)	2015
*created_at_year**	Year a project was created at (2009 - 2017)	2014	1.37	2015 (2014, 2015)	2015
launched_at_year	Year a project was launched at (2009 - 2017)	2015	1.37	2015 (2014, 2016)	2015
*deadline_month**	Month of deadline (1 - 12)	6.71	3.39	7 (4, 10)	8
*created_at_month**	Month a project was created at (1 - 12)	6.42	3.33	7 (4, 9)	7
launched_at_month	Month a project was launched at (1 - 12)	6.50	3.36	7 (4, 9)	7
deadline_day	Day of deadline (1 - 31)	15.66	9.06	15 (8, 23)	1
*created_at_day**	Day a project was created at (1 - 31)	15.54	8.79	15 (8, 23)	13
launched_at_day	Day a project was launched at (1 - 31)	15.27	8.83	15 (8, 23)	1
*deadline_weekday**	Weekday of deadline	Monday to Sunday (Details in appendix).			Friday
*created_at_weekday**	Weekday a project was created at				Tuesday
launched_at_weekday	Weekday a project was launched at				Tuesday

Table 1: Distribution Metrics. *(Variables with * were used in our Logistic model construction.)*

The results of our Logistic Regression model are shown in Table 2. For clarity, only significant predictors at the significance level of 0.05 in the model are shown, and results for non-significant predictors can be checked in appendix.

Variable	**Exp Rate (95% CI)**	**P-value**
Funding goal (1k USD)	0.989 (0.989, 0.989)	< 0.001
Length of project name (3 to 15 words)	1.106 (1.105, 1.106)	< 0.001
Month of deadline (1 to 12)	1.021 (1.021, 1.021)	0.005
Weekday of deadline: Sunday vs. Monday	0.813 (0.811, 0.815)	0.037
Year the project was created at (2009 to 2017)	0.830 (0.829, 0.830)	< 0.001
Month the project was created at (1 to 12)	0.984 (0.984, 0.984)	0.041
Weekday the project was created at: Saturday vs. Monday	0.747 (0.746, 0.749)	0.003
Weekday the project was created at: Sunday vs. Monday	0.796 (0.795, 0.798)	0.018
Number of days from launching a project to deadline	0.987 (0.987, 0.987)	< 0.001

Table 2: Coefficient Estimates for Significant Predictors (0.05 Significant Level) in Logistic Regression Model.

The model indicated that funding goal, length of project name, month and weekday of deadline, year, month, and weekday of creating a project, and the number of days from launching a project to deadline were significant contributors to the probability of success for a campaign on Kickstarter.

It can concluded from Table 2 that, if control values of other variables not change, at the significance level of 0.05, (1) the probability of a campaign being successful would decrease by 2.1% for every $1000 increase for funding goal; (2) 1 word increase for length of project name would increase the probability of a campaign being successful by 10.6%; (3) 1 month increase for deadline would increase the probability by 2.1%; (4) compared with Monday, setting Sunday as deadline would decrease the probability by 18.7%; (5) 1 year increase for creating time of a project would decrease the probability of a campaign being successful by 17%; (6) 1 month increase for creating time of project would decrease the probability by 2.6%; (7) compared with Monday, creating the project at Saturday and Sunday would crease the probability by 25.3% and 20.4% separately; (8) 1 day increase for timespan from launching the project to deadline would decrease the probability of a campaign being successful by 1.3%.

To predict the probability of success for the client's campaign, as explained in Method section, for significant predictors, we set the funding goal as $25,000, length of project name as 15 (max name length of all projects having raised over $25,000), deadline month as 12, year a project was created at as 2022, month a project was created at as 1, the number of days from launching to deadline as 1; for non-significant predictors, we set the length of blurb as 13 (median blurb length of all projects having raised over $25,000), disabled communication as FALSE (class accounting higher proportion), deadline day as 1 (mode value of all projects having raised over $25,000), day a project was created at as 22 (mode value of all projects having raised over $25,000), the number of days from creating to launching as 24 (median value of all projects having raised over $25,000), and project category as "unknown" (value of category for missing values in data set). In particular, for weekday of deadline, only Sunday decreased the probability of success significantly but other weekdays are non-significant, so we set the weekday of deadline as Friday (mode value of all projects having raised over $25,000); similarly, for weekday a project was created at, only Saturday and Sunday decreased the probability of success significantly but other weekdays are non-significant, so we set the weekday the project was created as Tuesday (mode value of all projects having raised over $25,000). Finally, our Logistic Regression mode predicted the probability of success for the client's campaign as 0.665.

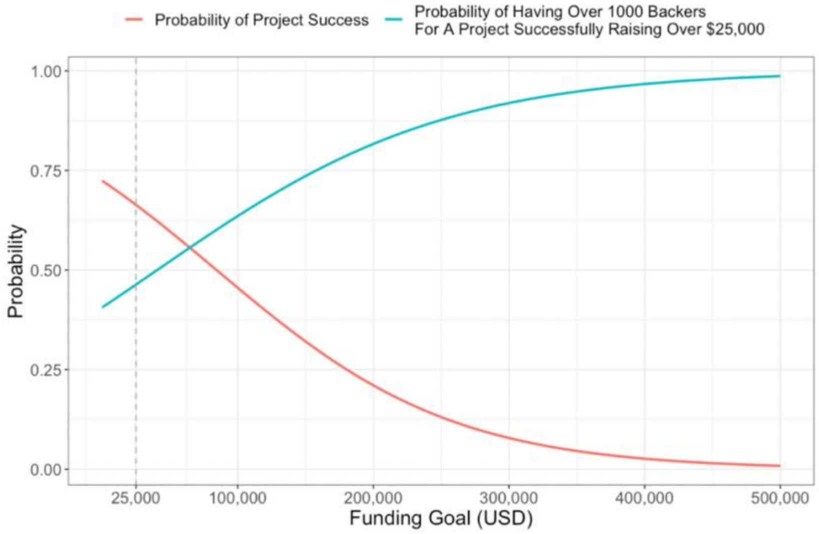


Figure 1: Trend of Probability of Campaign Success and Probability of Having at Least 1000 Backers for a Project Successfully Raising at Least $25,000 over Funding Goal.

The details on simple Logistic Regression model for predicting the probability of having over 1000 backers for a successful campaign having raised over $25,000 can be found in appendix. The same predictor values as above predicted the probability having over 1000 backers for a successful campaign having raised $25,000 as 0.463 which is not high. It indicated that it was not very likely that for a campaign raising $25,000 to have over 1000 backers.

Figure 1 shows the trade-off between probability of a campaign being successful and probability of having over 1000 backers after the campaign successfully raised the funding goal. The higher the funding goal is, the more likely the number of backers exceeds 1000 if the campaign succeeds. However, the higher the funding goal is, the less likely the campaign succeeds. Based on the model, we did not think it was necessary to set having at least 1000 backers as the part of the client's goal. On the one hand, the creator can only set the funding goal but not control the number of backers. On the other hand, higher funding goals are more likely to have more backers but less likely to be successfully raised. It is not necessary to set a higher funding goal to have a larger number of backers, which may sacrifice the probability of success for the campaign.

Conclusion

In conclusion, we accomplished the goals of determine generally what makes for a successful campaign using past Kickstarter campaign data and predicting the chance of success for their new community-based business aiming to raise at least $25,000.

Generally speaking, to increase the probability of a campaign being successful or to be more likely to succeed, the funding goal should be lower, the length of project name should be longer, the deadline month should be later in the year, the year and month of creating date for the campaign should be earlier, the deadline weekday should not be Sunday, the weekday the campaign is created at should not be Sunday and Saturday, and finally, the timespan from launching the project to deadline should be shorter.

Finally, based on the Logistic Regression model, the probability of success for the client's new community-based business aiming to raise $25,000 was predicted as 0.665 with imputation of other predictor values in the model. The probability of having at least 1000 backers for a project having successfully raised at least $25,000 was predicted as 0.463 which is not high. It was not recommended for the client to set the number of backers as part of the goal because the creator on Kickstarter cannot control the number of backers and there was a trade-off between probability of having at least 1000 backers and the probability of success for the campaign.

One possible limitation was that our analyses did not include text analysis on project name and blurb. Due to skill limit on text analysis, only length of project name and length of blurb were included in the model to provide some information on name and blurb. However, some keywords in the project name and blurb can have large impact for people on Kickstarter on deciding whether to pledge for a project. Text analysis can be taken into consideration in the future analysis.

STATS 504 Assignment 3 Kickstarter Success Appendix

In [1]:

```
# load libraries
library(tidyverse)
library(ggplot2)
library(corrplot)
```

```
── Attaching packages ───────────────────────────────────────────────────────────
─────────────────────────────────────────────────────── tidyverse 1.3.1 ──


✔ ggplot2 3.3.6     ✔ purrr   0.3.4
✔ tibble  3.1.7     ✔ dplyr   1.0.9
✔ tidyr   1.2.0     ✔ stringr 1.4.0
✔ readr   2.1.2     ✔ forcats 0.5.1

── Conflicts ────────────────────────────────────────────────────────────────────
──────────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()

corrplot 0.92 loaded
```

In [2]:

```
# load data
df <- read.csv("https://query.data.world/s/lxnrwj5w73bsigranne42td54f54sm", hea
```

In [3]:

```
head(df)
```

	X	id	
	<int>	<int>	

1	0	1454391034	w=160&h=90&fit=fill&bg=000000&v=1463719439&auto=format&q=92&s=362
			w=40&h=22&fit=fill&bg=000000&v=1463719439&auto=format&q=92&s=8c326
			w=1024&h=576&fit=fill&bg=000000&v=1463719439&auto=format&q=92&s=
			w=266&h=150&fit=fill&bg=000000&v=1463719439&auto=format&q=92&
			ugc.imgix.net/assets/011/959/953/4e53aa51f82e9764b135307761da1cde_or
			ugc.imgix.net/assets/011/959/953/4e53aa51f82e9764b135307761da1cde_o
			ugc.imgix.net/assets/011/959/953/4e53aa51f82e9764b135307761da1cde_or
			ugc.imgix.net/assets/011/959/953/4e53aa51f82e9764b135307761da1cde_original

2	1	1655206086	w=160&h=90&fit=fill&bg=000000&v=1463726814&auto=format&q=92&s=570
			w=40&h=22&fit=fill&bg=000000&v=1463726814&auto=format&q=92&s=5a6c14
			w=1024&h=576&fit=fill&bg=000000&v=1463726814&auto=format&q=92&s=a6
			w=266&h=150&fit=fill&bg=000000&v=1463726814&auto=format&q=92&s=b
			ugc.imgix.net/assets/012/043/791/0b63de0aa160746c6f26a0eed0ae6828_o
			ugc.imgix.net/assets/012/043/791/0b63de0aa160746c6f26a0eed0ae6828_c
			ugc.imgix.net/assets/012/043/791/0b63de0aa160746c6f26a0eed0ae6828_orig
			ugc.imgix.net/assets/012/043/791/0b63de0aa160746c6f26a0eed0ae6828_origina

3	2	311581827	w=160&h=90&fit=fill&bg=000000&v=1463723952&auto=format&q=92&s=613e
			w=40&h=22&fit=fill&bg=000000&v=1463723952&auto=format&q=92&s=c6dd0e
			w=1024&h=576&fit=fill&bg=000000&v=1463723952&auto=format&q=92&s=
			w=266&h=150&fit=fill&bg=000000&v=1463723952&auto=format&q=92&s=
			ugc.imgix.net/assets/012/012/056/c566aeb9b51df01e8dd2828ce97d753f_ori
			ugc.imgix.net/assets/012/012/056/c566aeb9b51df01e8dd2828ce97d753f_ori
			ugc.imgix.net/assets/012/012/056/c566aeb9b51df01e8dd2828ce97d753f_c
			ugc.imgix.net/assets/012/012/056/c566aeb9b51df01e8dd2828ce97d753f_origina

4	3	859724515	w=160&h=90&fit=fill&bg=000000&v=1463705583&auto=format&q=92&s=fe33
			w=40&h=22&fit=fill&bg=000000&v=1463705583&auto=format&q=92&s=ea2bb
			w=1024&h=576&fit=fill&bg=000000&v=1463705583&auto=format&q=92&s=29
			w=266&h=150&fit=fill&bg=000000&v=1463705583&auto=format&q=92&s=94e
			ugc.imgix.net/assets/011/860/879/620804a20f84c31d4f53a80313635842_origi
			ugc.imgix.net/assets/011/860/879/620804a20f84c31d4f53a80313635842_or
			ugc.imgix.net/assets/011/860/879/620804a20f84c31d4f53a80313635842_orig
			ugc.imgix.net/assets/011/860/879/620804a20f84c31d4f53a80313635842_origin

5	4	1613604977	w= w=266&h=150&fit=fill&bg=FFFFFF&v=1464815065&auto=format&frame=1&q=92 ugc.imgix.net/assets/012/521/917/305ee995fe695b1920f5e415f12faa15_origina ugc.imgix.net/assets/012/521/917/305ee995fe695b1920f5e415

	X	id	
	<int>	<int>	

6	5	808486483	w=160&h=90&fit=fill&bg=000000&v=1463750513&auto=format&q=92&s=520
			w=40&h=22&fit=fill&bg=000000&v=1463750513&auto=format&q=92&s=81b9f78
			w=1024&h=576&fit=fill&bg=000000&v=1463750513&auto=format&q=92&s=
			w=266&h=150&fit=fill&bg=000000&v=1463750513&auto=format&q=92&s=e
			ugc.imgix.net/assets/012/283/666/4dc7472c8cb40252e48ee1dbcd7097eb_ori
			ugc.imgix.net/assets/012/283/666/4dc7472c8cb40252e48ee1dbcd7097eb_or
			ugc.imgix.net/assets/012/283/666/4dc7472c8cb40252e48ee1dbcd7097eb_ori
			ugc.imgix.net/assets/012/283/666/4dc7472c8cb40252e48ee1dbcd7097eb_origin

```
In [4]: colnames(df)
```

'X' · 'id' · 'photo' · 'name' · 'blurb' · 'goal' · 'pledged' · 'state' · 'slug' · 'disable_communication' · 'country' · 'currency' · 'currency_symbol' · 'currency_trailing_code' · 'deadline' · 'state_changed_at' · 'created_at' · 'launched_at' · 'staff_pick' · 'backers_count' · 'static_usd_rate' · 'usd_pledged' · 'creator' · 'location' · 'category' · 'profile' · 'spotlight' · 'urls' · 'source_url' · 'friends' · 'is_starred' · 'is_backing' · 'permissions' · 'name_len' · 'name_len_clean' · 'blurb_len' · 'blurb_len_clean' · 'deadline_weekday' · 'state_changed_at_weekday' · 'created_at_weekday' · 'launched_at_weekday' · 'deadline_month' · 'deadline_day' · 'deadline_yr' · 'deadline_hr' · 'state_changed_at_month' · 'state_changed_at_day' · 'state_changed_at_yr' · 'state_changed_at_hr' · 'created_at_month' · 'created_at_day' · 'created_at_yr' · 'created_at_hr' · 'launched_at_month' · 'launched_at_day' · 'launched_at_yr' · 'launched_at_hr' · 'create_to_launch' · 'launch_to_deadline' · 'launch_to_state_change' · 'create_to_launch_days' · 'launch_to_deadline_days' · 'launch_to_state_change_days' · 'SuccessfulBool' · 'USorGB' · 'TOPCOUNTRY' · 'LaunchedTuesday' · 'DeadlineWeekend'

```
In [8]: cor(as.numeric(strptime(df$state_changed_at, '%Y-%m-%d %H:%M:%S')), as.numeric(
```

```
0.999740329853206
```

```
In [9]: length(unique(df$blurb))
```

```
20462
```

```
In [10]: length(unique(df$name))
```

```
20611
```

```
In [11]: nrow(df)
```

```
20632
```

- do not use blurb: almost every project has its own blurb; meaningless to use it for prediction <- use blurb length instead
- do not use name: almost every project has its own name; meaningless to use it for prediction <- use name length instead

In [13]:

```
# select important variables
data = df %>% transmute(
    project = id,
    funding_goal = goal,
    name = name,
    name_len = name_len,
    blurb_len = blurb_len_clean,
    pledged = pledged,
    backers = backers_count,
    state = factor(state),
    success = SuccessfulBool,
    disable_communication = as.factor(disable_communication),
    deadline_year = as.integer(deadline_yr),
    deadline_month = as.integer(deadline_month),
    deadline_day = as.integer(deadline_day),
    deadline_weekday = as.factor(deadline_weekday),
    created_at_year = as.integer(created_at_yr),
    created_at_month = as.integer(created_at_month),
    created_at_day = as.integer(created_at_day),
    created_at_weekday = as.factor(created_at_weekday),
    launched_at_year = as.integer(launched_at_yr),
    launched_at_month = as.integer(launched_at_month),
    launched_at_day = as.integer(launched_at_day),
    launched_at_weekday = as.factor(launched_at_weekday),
    create2launch = create_to_launch_days,
    launch2deadline = launch_to_deadline_days,
    country = factor(country),
    currency = factor(currency),
    category = factor(ifelse(category=='', 'Unknown', category))
)
```

In [14]:

```
rm(df)
```

In [15]:

```
USdata = data %>% filter(country=='US' & currency == 'USD') %>% select(-c(count
```

In [16]:

```
rm(data)
```

In [17]:

```
head(USdata)
```

	project	funding_goal	name	name_len	blurb_len	pledged	backers	state	suc
	<int>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<int>	<fct>	<
1	1454391034	1500	Auntie Di's Music Time Sign ASL for Hearing and HOH Children	11	16	0	0	failed	
2	1655206086	500	Jump Start Kindergarten Toolkit	4	15	0	0	failed	
3	311581827	100000	Ojukwu Balewa Awolowo (O.B.A.) Public Library Of Nigeria	8	10	120	5	failed	
4	859724515	5000	MASTIZE - [mas-TAHYZ, MAS-tahyz] - to spread	7	13	0	0	failed	
5	808486483	13000	Shadow School Board - Reforming Texas School Boards	8	15	1136	12	failed	
6	883246296	50000	Research in HIV prevention, treatment, and aid	7	13	0	0	failed	

Exploratory Data Analysis

```
In [18]: str(USdata)
```

```
'data.frame':   14141 obs. of  25 variables:
 $ project              : int  1454391034 1655206086 311581827 859724515 80848
6483 883246296 242834615 1624645868 429226406 1849446483 ...
 $ funding_goal         : num  1500 500 100000 5000 13000 50000 10000 15000 10
000 10000 ...
 $ name                 : chr  "Auntie Di's Music Time Sign ASL for Hearing an
d HOH Children" "Jump Start Kindergarten Toolkit" "Ojukwu Balewa Awolowo (O.B.
A.) Public Library Of Nigeria" "MASTIZE - [mas-TAHYZ, MAS-tahyz]  - to spread"
...
 $ name_len             : num  11 4 8 7 8 7 3 5 6 5 ...
 $ blurb_len            : num  16 15 10 13 15 13 12 13 13 13 ...
 $ pledged              : num  0 0 120 0 1136 ...
 $ backers              : int  0 0 5 0 12 0 0 0 10 7 ...
 $ state                : Factor w/ 5 levels "canceled","failed",..: 2 2 2 2 2
2 2 2 2 2 ...
 $ success              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ disable_communication: Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1
1 ...
 $ deadline_year        : int  2015 2015 2015 2014 2015 2015 2015 2015 2016 20
15 ...
 $ deadline_month       : int  1 5 3 10 11 5 9 12 6 6 ...
 $ deadline_day         : int  23 1 26 6 20 29 27 2 30 1 ...
 $ deadline_weekday     : Factor w/ 7 levels "Friday","Monday",..: 1 1 5 2 1 1
4 7 5 2 ...
 $ created_at_year      : int  2014 2015 2015 2014 2015 2015 2015 2015 2016 20
15 ...
 $ created_at_month     : int  11 2 1 9 10 4 8 11 3 3 ...
 $ created_at_day       : int  29 20 24 5 19 29 13 1 22 20 ...
 $ created_at_weekday   : Factor w/ 7 levels "Friday","Monday",..: 3 1 3 1 2 7
5 4 6 1 ...
 $ launched_at_year     : int  2014 2015 2015 2014 2015 2015 2015 2015 2016 20
15 ...
 $ launched_at_month    : int  12 3 1 9 10 4 8 11 5 4 ...
 $ launched_at_day      : int  17 2 25 6 21 29 13 2 1 17 ...
 $ launched_at_weekday  : Factor w/ 7 levels "Friday","Monday",..: 7 2 4 3 7 7
5 2 4 1 ...
 $ create2launch        : int  17 10 1 0 2 0 0 1 39 28 ...
 $ launch2deadline      : int  36 60 60 30 30 30 45 30 60 45 ...
 $ category             : Factor w/ 25 levels "Academic","Apps",..: 1 1 1 1 1
1 1 1 1 1 ...
```

In [19]:

```
summary(USdata)
```

```
   project           funding_goal          name              name_len
Min.   :2.610e+05   Min.   :        1   Length:14141       Min.   : 1.000
1st Qu.:5.495e+08   1st Qu.:     5000   Class :character   1st Qu.: 4.000
Median :1.071e+09   Median :    15000   Mode  :character   Median : 6.000
Mean   :1.073e+09   Mean   :    88661                      Mean   : 5.998
3rd Qu.:1.606e+09   3rd Qu.:    50000                      3rd Qu.: 8.000
Max.   :2.147e+09   Max.   :100000000                      Max.   :16.000
                                                           NA's   :3
  blurb_len        pledged            backers                state
Min.   : 1.00   Min.   :      0   Min.   :     0.0   canceled  :1663
1st Qu.:11.00   1st Qu.:     37   1st Qu.:     2.0   failed    :7668
Median :13.00   Median :    929   Median :    14.0   live      : 306
Mean   :13.02   Mean   :  24947   Mean   :   216.6   successful:4362
3rd Qu.:15.00   3rd Qu.:   7381   3rd Qu.:    75.0   suspended : 142
Max.   :30.00   Max.   :6225355   Max.   :105857.0
NA's   :3
   success       disable_communication deadline_year  deadline_month
Min.   :0.0000   False:13999           Min.   :2009   Min.   : 1.000
1st Qu.:0.0000   True :  142           1st Qu.:2014   1st Qu.: 4.000
Median :0.0000                         Median :2015   Median : 7.000
Mean   :0.3085                         Mean   :2015   Mean   : 6.716
3rd Qu.:1.0000                         3rd Qu.:2016   3rd Qu.:10.000
Max.   :1.0000                         Max.   :2017   Max.   :12.000

 deadline_day    deadline_weekday created_at_year created_at_month
Min.   : 1.00   Friday   :2594    Min.   :2009    Min.   : 1.000
1st Qu.: 8.00   Monday   :1504    1st Qu.:2014    1st Qu.: 4.000
Median :15.00   Saturday :2015    Median :2015    Median : 7.000
Mean   :15.65   Sunday   :2074    Mean   :2014    Mean   : 6.415
3rd Qu.:23.00   Thursday :2376    3rd Qu.:2015    3rd Qu.: 9.000
Max.   :31.00   Tuesday  :1364    Max.   :2017    Max.   :12.000
                Wednesday:2214
created_at_day  created_at_weekday launched_at_year launched_at_month
Min.   : 1.00   Friday   :1863     Min.   :2009     Min.   : 1.00
1st Qu.: 8.00   Monday   :2365     1st Qu.:2014     1st Qu.: 4.00
Median :15.00   Saturday :1430     Median :2015     Median : 7.00
Mean   :15.55   Sunday   :1533     Mean   :2015     Mean   : 6.51
3rd Qu.:23.00   Thursday :2152     3rd Qu.:2016     3rd Qu.: 9.00
Max.   :31.00   Tuesday  :2481     Max.   :2017     Max.   :12.00
                Wednesday:2317
launched_at_day launched_at_weekday create2launch     launch2deadline
Min.   : 1.00   Friday   :1958      Min.   :   0.00   Min.   : 1.0
1st Qu.: 8.00   Monday   :2902      1st Qu.:   4.00   1st Qu.:30.0
Median :15.00   Saturday : 724      Median :  15.00   Median :30.0
Mean   :15.28   Sunday   : 697      Mean   :  53.78   Mean   :34.8
3rd Qu.:23.00   Thursday :2096      3rd Qu.:  49.00   3rd Qu.:40.0
Max.   :31.00   Tuesday  :3223      Max.   :1754.00   Max.   :91.0
                Wednesday:2541
    category
Hardware:2448
Web     :2016
Software:1828
Gadgets :1667
Unknown :1336
Plays   : 765
(Other) :4081
```

In [20]:

```
nrow(USdata)
```

```
14141
```

In [21]:
```
# remove 3 NA's for blurb_len and name_len
USdata = na.omit(USdata)
nrow(USdata)
```

14138

In [22]:
```
USdata %>% group_by(state, success) %>% summarize(count = n()/nrow(USdata)*100)
```

```
`summarise()` has grouped output by 'state'. You can override using the `.grou
ps` argument.
```

A grouped_df: 5 × 3

state	success	count
<fct>	<int>	<dbl>
canceled	0	11.741406
failed	0	54.236809
live	0	2.164380
successful	1	30.853020
suspended	0	1.004385

We'll user successBool as the response.

In [23]:
```
colnames(USdata)
```

'project' · 'funding_goal' · 'name' · 'name_len' · 'blurb_len' · 'pledged' · 'backers' · 'state' · 'success' · 'disable_communication' · 'deadline_year' · 'deadline_month' · 'deadline_day' · 'deadline_weekday' · 'created_at_year' · 'created_at_month' · 'created_at_day' · 'created_at_weekday' · 'launched_at_year' · 'launched_at_month' · 'launched_at_day' · 'launched_at_weekday' · 'create2launch' · 'launch2deadline' · 'category'

In [24]:
```
cat.vars = c('category', 'disable_communication', 'deadline_weekday', 'created_
date.vars = c('deadline_year', 'deadline_month', 'deadline_day',
              'created_at_year', 'created_at_month', 'created_at_day',
              'launched_at_year', 'launched_at_month', 'launched_at_day')
num.vars = colnames(USdata)[which(!colnames(USdata) %in% c(cat.vars, date.vars,
```

In [25]:
```
num.hist = USdata[, c(num.vars, 'success')] %>% gather(key = "variable", value
```

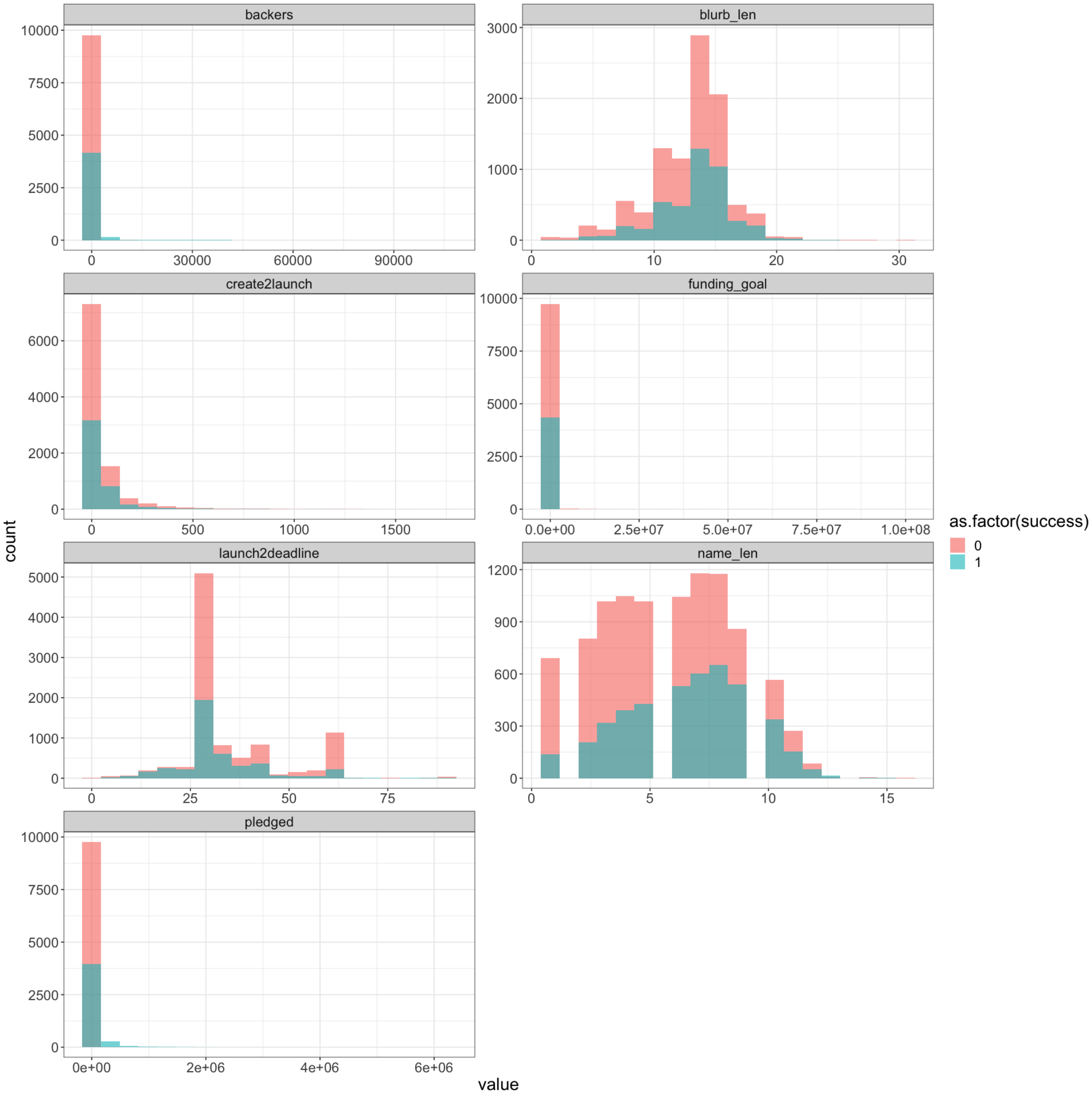
In [26]:
```
head(num.hist)
```

A data.frame: 6 × 3

	success	variable	value
	<int>	<chr>	<dbl>
1	0	funding_goal	1500
2	0	funding_goal	500
3	0	funding_goal	100000
4	0	funding_goal	5000
5	0	funding_goal	13000
6	0	funding_goal	50000

In [27]:

```
options(repr.plot.width = 16, repr.plot.height = 16)
# histogram for numerical variables
num.hist %>% ggplot() +
    geom_histogram(aes(x = value, fill = as.factor(success)), bins = 20, alpha=
    facet_wrap(~variable, scales = 'free', ncol = 2) + theme_bw() +
    theme(text = element_text(size = 18))
```



There exists outliers for each numerical variables except for blurb_len, name_len, create2launch, launch2deadline.

In [28]:

```
# filter outliers for backers
USdata %>% filter(backers > 10000) %>% group_by(success) %>% summarize(count =
```

A tibble: 1 × 2

success	count
<int>	<int>
1	29

In [29]:

```
# filter outliers for create2lauch
USdata %>% filter(create2launch > 500) %>% group_by(success) %>% summarize(cour
```

A tibble: 2 × 2

success	count
<int>	<int>
0	173
1	62

In [30]:
```
# filter outliers for funding goal
USdata %>% filter(funding_goal > 5e+5) %>% group_by(success) %>% summarize(cour
```

A tibble: 2 × 2

success	count
<int>	<int>
0	171
1	5

In [31]:
```
# filter outliers for pledged
# may not be used to construct model
USdata %>% filter(pledged > 5e+5) %>% group_by(success) %>% summarize(count = r
```

A tibble: 2 × 2

success	count
<int>	<int>
0	6
1	132

In [32]:
```
US.filtered = USdata %>% filter(backers <= 10000 & create2launch <= 500 & fundi
                                   & pledged <= 5e+5)
```

In [111…
```
nrow(US.filtered)
```

13598

In [112…
```
# number of rows removed
14138 - 13598
```

540

In [114…
```
length(unique(US.filtered$project))
```

13598

In [117…
```
length(unique(US.filtered$name))
```

13590

In [119…
```
US.filtered %>% group_by(name) %>% summarise(count = n()) %>% arrange(desc(cour
```

A tibble: 10 × 2

name	count
<chr>	<int>
BEIRUT, LADY OF LEBANON	2
Cancelled. (Canceled)	2
FREE ENERGY	2
Gruesome Playground Injuries	2
Project Canceled (Canceled)	2
test (Canceled)	2
Us, Bent (Canceled)	2
weSTAND: A Stand With a Mission	2
¡Latin Food Fest! Mobile App and Magazine	1
¡OSO FABULOSO & The Bear Backs!	1

In [121…

```
table(US.filtered$success)
```

```
   0    1 
9430 4168 
```

In [33]:

```
summary(US.filtered)
```

```
   project           funding_goal        name              name_len     
Min.   :2.610e+05   Min.   :     1   Length:13598       Min.   : 1.000  
1st Qu.:5.491e+08   1st Qu.:  4500   Class :character   1st Qu.: 4.000  
Median :1.073e+09   Median : 13056   Mode  :character   Median : 6.000  
Mean   :1.073e+09   Mean   : 35911                      Mean   : 5.985  
3rd Qu.:1.606e+09   3rd Qu.: 42000                      3rd Qu.: 8.000  
Max.   :2.147e+09   Max.   :500000                      Max.   :16.000  

  blurb_len        pledged          backers                state     
Min.   : 1.00   Min.   :     0   Min.   :   0.0   canceled  :1591  
1st Qu.:11.00   1st Qu.:    37   1st Qu.:   2.0   failed    :7408  
Median :13.00   Median :   911   Median :  14.0   live      : 295  
Mean   :13.01   Mean   : 15014   Mean   : 142.5   successful:4168  
3rd Qu.:15.00   3rd Qu.:  7016   3rd Qu.:  72.0   suspended : 136  
Max.   :30.00   Max.   :492204   Max.   :9895.0                    

   success       disable_communication deadline_year  deadline_month 
Min.   :0.0000   False:13462           Min.   :2009   Min.   : 1.00  
1st Qu.:0.0000   True :  136           1st Qu.:2014   1st Qu.: 4.00  
Median :0.0000                         Median :2015   Median : 7.00  
Mean   :0.3065                         Mean   :2015   Mean   : 6.71  
3rd Qu.:1.0000                         3rd Qu.:2016   3rd Qu.:10.00  
Max.   :1.0000                         Max.   :2017   Max.   :12.00  

 deadline_day    deadline_weekday created_at_year created_at_month
Min.   : 1.00   Friday   :2480    Min.   :2009    Min.   : 1.000  
1st Qu.: 8.00   Monday   :1451    1st Qu.:2014    1st Qu.: 4.000  
Median :15.00   Saturday :1941    Median :2015    Median : 7.000  
Mean   :15.66   Sunday   :2007    Mean   :2014    Mean   : 6.417  
3rd Qu.:23.00   Thursday :2275    3rd Qu.:2015    3rd Qu.: 9.000  
Max.   :31.00   Tuesday  :1308    Max.   :2017    Max.   :12.000  
                Wednesday:2136                                    
created_at_day  created_at_weekday launched_at_year launched_at_month
Min.   : 1.00   Friday   :1787     Min.   :2009     Min.   : 1.000   
1st Qu.: 8.00   Monday   :2292     1st Qu.:2014     1st Qu.: 4.000   
Median :15.00   Saturday :1381     Median :2015     Median : 7.000   
Mean   :15.54   Sunday   :1474     Mean   :2015     Mean   : 6.502   
3rd Qu.:23.00   Thursday :2061     3rd Qu.:2016     3rd Qu.: 9.000   
Max.   :31.00   Tuesday  :2381     Max.   :2017     Max.   :12.000   
                Wednesday:2222                                       
launched_at_day launched_at_weekday create2launch    launch2deadline
Min.   : 1.00   Friday   :1897      Min.   :  0.00   Min.   : 1.00  
1st Qu.: 8.00   Monday   :2789      1st Qu.:  4.00   1st Qu.:30.00  
Median :15.00   Saturday : 700      Median : 14.00   Median :30.00  
Mean   :15.27   Sunday   : 675      Mean   : 41.69   Mean   :34.66  
3rd Qu.:23.00   Thursday :2026      3rd Qu.: 45.00   3rd Qu.:40.00  
Max.   :31.00   Tuesday  :3082      Max.   :495.00   Max.   :91.00  
                Wednesday:2429                                      
    category   
Hardware:2315  
Web     :1958  
Software:1759  
Gadgets :1601  
Unknown :1269  
Plays   : 757  
(Other) :3939  
```

In [302…

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
```

```
}
tmp = c(date.vars, 'deadline_weekday', 'created_at_weekday', 'launched_at_week
for (v in tmp) {
    print(paste0(v, ': ', getmode(US.filtered[, v])))
}
```

```
[1] "deadline_year: 2015"
[1] "deadline_month: 8"
[1] "deadline_day: 1"
[1] "created_at_year: 2015"
[1] "created_at_month: 7"
[1] "created_at_day: 13"
[1] "launched_at_year: 2015"
[1] "launched_at_month: 7"
[1] "launched_at_day: 1"
[1] "deadline_weekday: Friday"
[1] "created_at_weekday: Tuesday"
[1] "launched_at_weekday: Tuesday"
```

In [301…

```
# std for date.vars
sqrt(diag(var(US.filtered[, date.vars])))
```

deadline_year: 1.37313678416855 **deadline_month:** 3.38811991765354 **deadline_day:** 9.06463391861006 **created_at_year:** 1.37160593667088 **created_at_month:** 3.33400605206473 **created_at_day:** 8.79039554938089 **launched_at_year:** 1.37289717953381 **launched_at_month:** 3.36191198336573 **launched_at_day:** 8.83293619883875

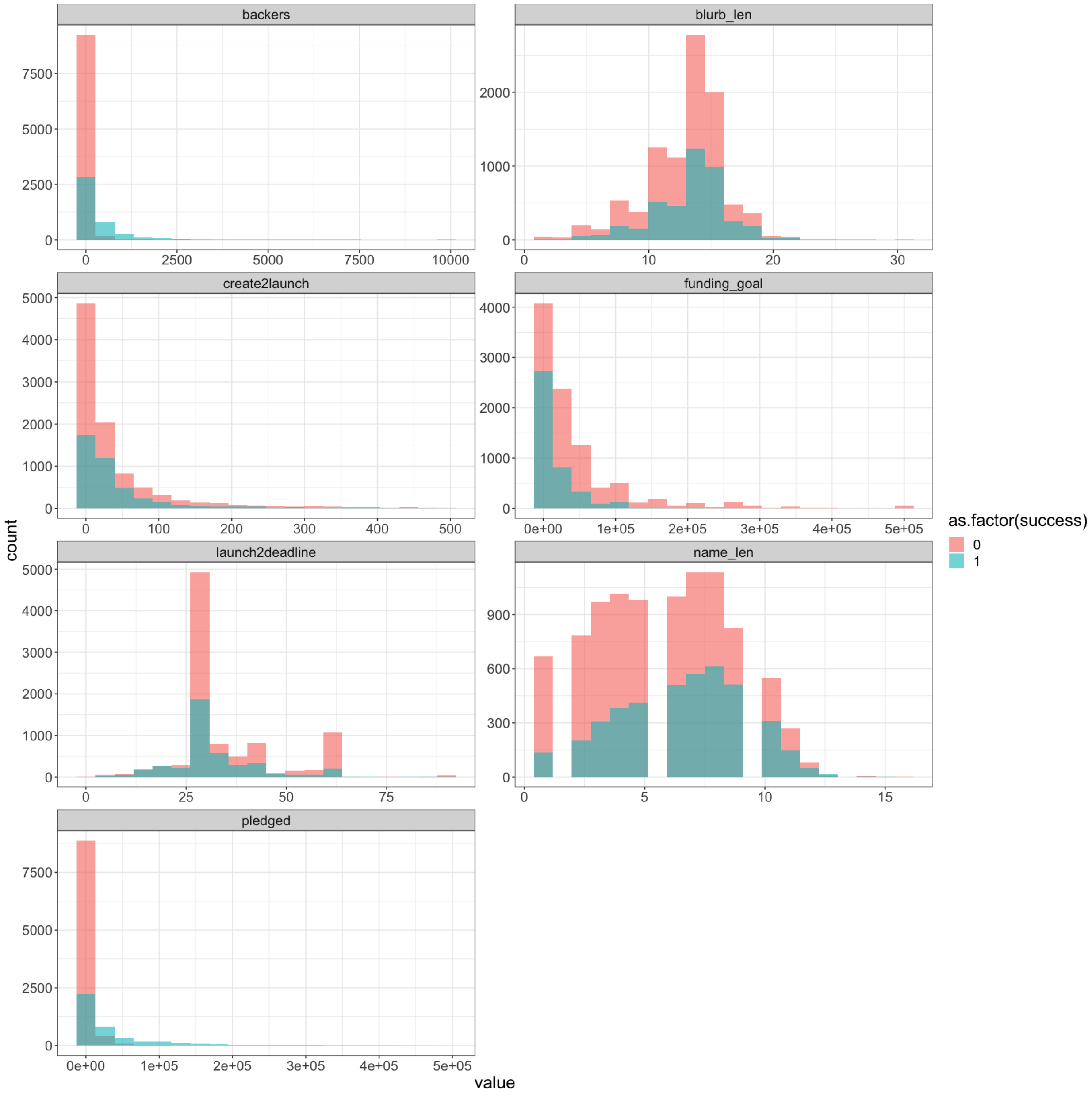
In [34]:

```
# std for num.vars
sqrt(diag(var(US.filtered[, num.vars])))
```

funding_goal: 62217.5951429982 **name_len:** 2.79141573926843 **blurb_len:** 3.20969638244238 **pledged:** 44901.9450242726 **backers:** 479.864878863694 **create2launch:** 70.5517744538496 **launch2deadline:** 11.9035564101078

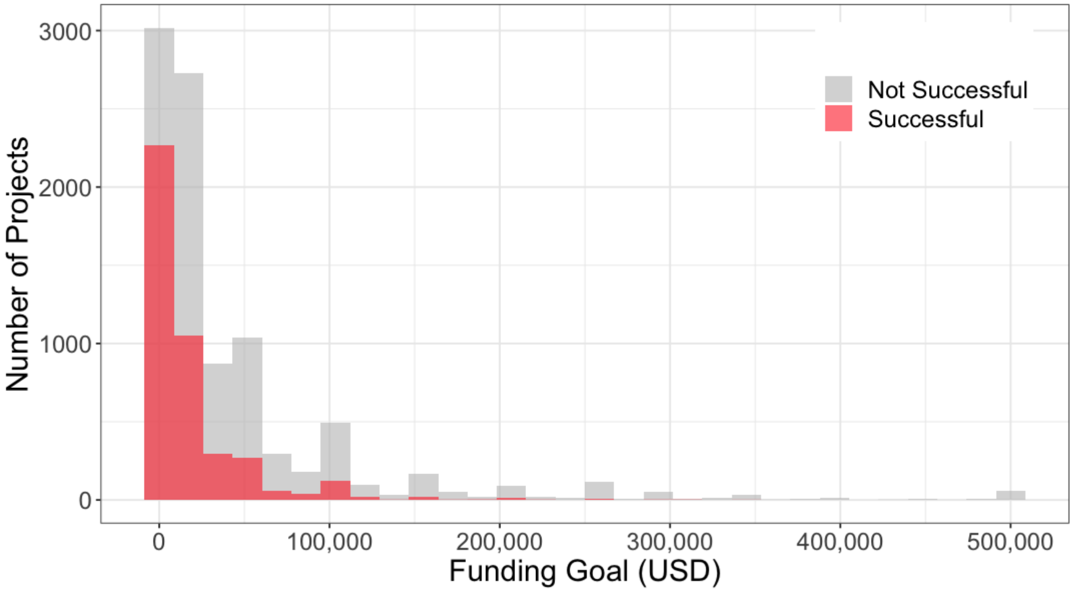
In [35]:

```
num.hist = US.filtered[, c(num.vars, 'success')] %>% gather(key = "variable", v
options(repr.plot.width = 16, repr.plot.height = 16)
# histogram for numerical variables
num.hist %>% ggplot() +
    geom_histogram(aes(x = value, fill = as.factor(success)), bins = 20, alpha=
    facet_wrap(~variable, scales = 'free', ncol = 2) + theme_bw() +
    theme(text = element_text(size = 18))
```



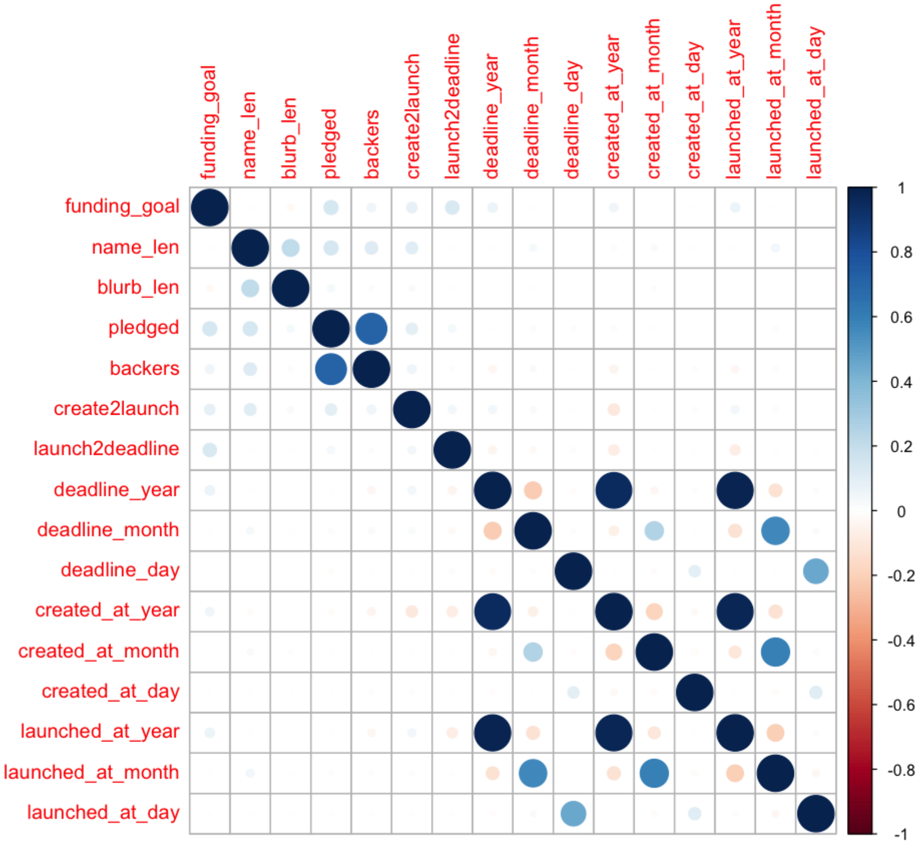
In [155…

```
options(repr.plot.width = 9, repr.plot.height = 5)
US.filtered %>% ggplot() +
    geom_histogram(aes(x = funding_goal, fill = as.factor(success)),
                   bins = 30, alpha=0.6, position = 'identity') +
    theme_bw() +
    theme(text = element_text(size = 18), legend.position = c(0.85, 0.85)) +
    scale_fill_manual(name = "", labels = c('Not Successful', 'Successful'),
                      values = c('gray', 'firebrick1')) +
    scale_x_continuous(breaks = c(0, 1e+5, 2e+5, 3e+5, 4e+5, 5e+5),
                       labels = c('0', '100,000', '200,000', '300,000', '400,00
    xlab('Funding Goal (USD)') + ylab('Number of Projects')
```



In [36]:

```
# correlation
options(repr.plot.width = 8, repr.plot.height = 8)
correlations <- cor(US.filtered[, c(num.vars, date.vars)])
corrplot(correlations, method="circle")
```

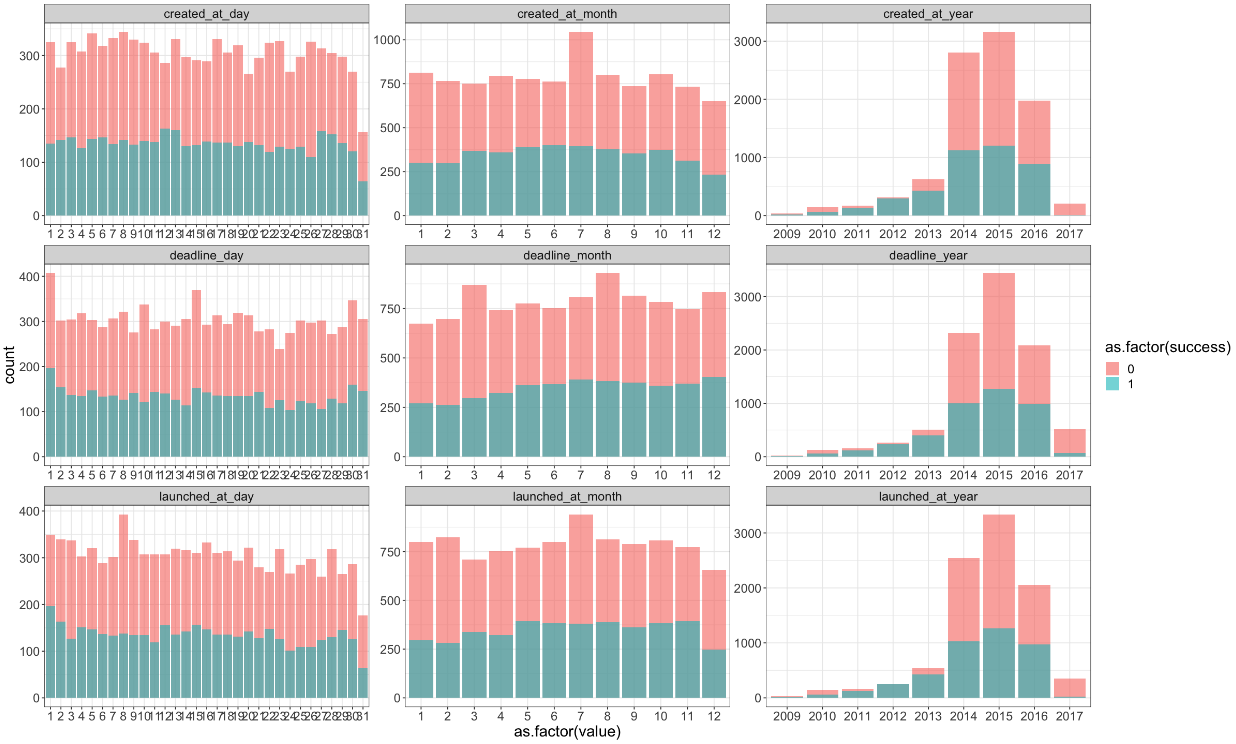


- Strong correlation between backers and pledged.

- Strong correlation between launched_at_month and deadline_month, created_at_month. <- only use created_month and deadline_month <- information of launched month can be covered in create2launch.

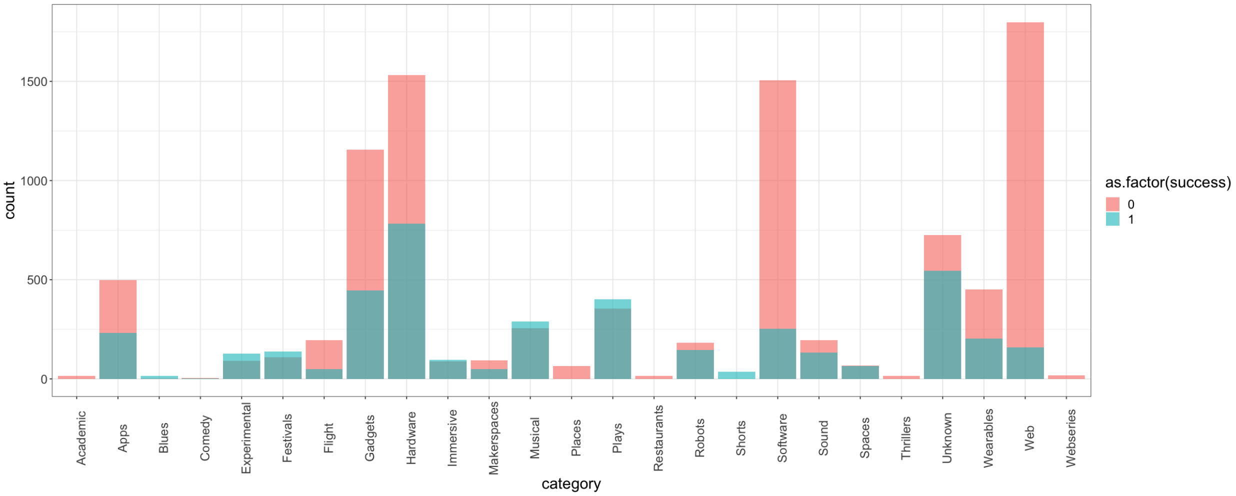
In [38]:

```
date.hist = US.filtered[, c(date.vars, 'success')] %>% gather(key = "variable",
options(repr.plot.width = 20, repr.plot.height = 12)
# histogram for date variables
date.hist %>% ggplot() +
    geom_bar(aes(x = as.factor(value), fill = as.factor(success)),alpha=0.6, pc
    facet_wrap(~variable, scales = 'free', ncol = 3) + theme_bw() +
    theme(text = element_text(size = 18))
```



In [39]:

```
options(repr.plot.width = 20, repr.plot.height = 8)
# categorical variable
US.filtered %>% select(category, success) %>% ggplot() +
    geom_bar(aes(x = category, fill = as.factor(success)), alpha = 0.6, positic
    theme_bw() +
    theme(text = element_text(size = 18), axis.text.x = element_text(angle = 9(
```



In [40]:

```
# calculate the proportion of success for each category
US.filtered %>% select(category, success) %>% group_by(category) %>%
    summarise(success_rate = sum(success)/n()) %>% arrange(desc(success_rate))
```

A tibble: 25 × 2

category	success_rate
<fct>	<dbl>
Shorts	1.00000000
Blues	0.88888889
Experimental	0.58715596
Festivals	0.56097561
Plays	0.53104359
Musical	0.53027523
Immersive	0.52150538
Spaces	0.49624060
Robots	0.44545455
Unknown	0.42947203
Sound	0.40366972
Makerspaces	0.34042553
Hardware	0.33779698
Apps	0.31917808
Wearables	0.31039755
Gadgets	0.27795128
Flight	0.20000000
Comedy	0.16666667
Software	0.14383172
Web	0.08120531
Academic	0.00000000
Places	0.00000000
Restaurants	0.00000000
Thrillers	0.00000000
Webseries	0.00000000

Modeling

In [41]:

```
table(US.filtered$success)
```

```
   0    1 
9430 4168 
```

Imbalanced class problem.

In [108…
```
9430/(9430+4168)
```

0.693484335931755

In [42]:
```
table(US.filtered$success[which(US.filtered$funding_goal > 25000)])
```

```
   0    1 
3693  855 
```

In [110…
```
855/(9430+4168)
```

0.0628768936608325

In [43]:
```
table(US.filtered$success[which(US.filtered$funding_goal <= 25000)])
```

```
   0    1 
5737 3313 
```

In [44]:
```
# reconstruct the data set by resampling the dataset
set.seed(1234)
num_samples = length(rownames(US.filtered)[which(US.filtered$success==1)])
sample_index = sample(rownames(US.filtered)[which(US.filtered$success==0)], nun
sample_index = sort(as.integer(sample_index))
```

In [45]:
```
US.sampled = rbind(US.filtered[which(US.filtered$success==1), ], US.filtered[sa
```

In [46]:
```
nrow(US.sampled)
```

8336

In [47]:
```
table(US.sampled$success)
```

```
   0    1 
4168 4168 
```

In [48]:
```
table(US.sampled$success[which(US.sampled$funding_goal > 25000)])
```

```
   0    1 
1613  855 
```

In [49]:
```
table(US.sampled$success[which(US.sampled$funding_goal <= 25000)])
```

```
   0    1 
2555 3313 
```

Logistic Regression

In [50]:
```
colnames(US.filtered)
```

'project' · 'funding_goal' · 'name' · 'name_len' · 'blurb_len' · 'pledged' · 'backers' · 'state' · 'success' · 'disable_communication' · 'deadline_year' · 'deadline_month' · 'deadline_day' · 'deadline_weekday' · 'created_at_year' · 'created_at_month' · 'created_at_day' · 'created_at_weekday' · 'launched_at_year' · 'launched_at_month' · 'launched_at_day' · 'launched_at_weekday' · 'create2launch' · 'launch2deadline' · 'category'

In [51]:

```
US.sampled$deadline_weekday = relevel(US.sampled$deadline_weekday, ref = 'Monda
US.sampled$created_at_weekday = relevel(US.sampled$created_at_weekday, ref = 'M
```

In [253…

```
US.sampled$funding_goal1000 = US.sampled$funding_goal/1000
```

In [254…

```
lr = glm(success ~ funding_goal1000 + name_len + blurb_len + disable_communicat
         deadline_month + deadline_day + deadline_weekday +
         created_at_year + created_at_month + created_at_day + created_at_weekc
         create2launch + launch2deadline + category,
         data = US.sampled,
         family = "binomial")
```

In [255…

```
summary(lr)
```

```
Call:
glm(formula = success ~ funding_goal1000 + name_len + blurb_len + 
    disable_communication + deadline_month + deadline_day + deadline_weekday + 
    created_at_year + created_at_month + created_at_day + created_at_weekday + 
    create2launch + launch2deadline + category, family = "binomial", 
    data = US.sampled)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.34000  -1.00522   0.00011   0.94757   2.58742  

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                  3.596e+02  1.158e+03   0.311  0.75613    
funding_goal1000            -1.143e-02  7.454e-04 -15.340  < 2e-16 ***
name_len                     1.005e-01  9.576e-03  10.492  < 2e-16 ***
blurb_len                    8.240e-03  8.006e-03   1.029  0.30341    
disable_communicationTrue   -1.650e+01  2.725e+02  -0.061  0.95171    
deadline_month               2.109e-02  7.544e-03   2.796  0.00518 ** 
deadline_day                -1.800e-03  2.730e-03  -0.659  0.50974    
deadline_weekdayFriday      -6.096e-03  9.491e-02  -0.064  0.94878    
deadline_weekdaySaturday    -9.810e-02  1.008e-01  -0.973  0.33036    
deadline_weekdaySunday      -2.072e-01  9.954e-02  -2.081  0.03742 *  
deadline_weekdayThursday     8.626e-02  9.684e-02   0.891  0.37306    
deadline_weekdayTuesday      5.836e-02  1.101e-01   0.530  0.59620    
deadline_weekdayWednesday   -8.379e-04  9.876e-02  -0.008  0.99323    
created_at_year             -1.865e-01  2.012e-02  -9.270  < 2e-16 ***
created_at_month            -1.611e-02  7.888e-03  -2.043  0.04108 *  
created_at_day              -2.100e-03  2.810e-03  -0.747  0.45491    
created_at_weekdayFriday    -1.381e-01  9.179e-02  -1.504  0.13250    
created_at_weekdaySaturday  -2.911e-01  9.886e-02  -2.945  0.00323 ** 
created_at_weekdaySunday    -2.279e-01  9.621e-02  -2.369  0.01783 *  
created_at_weekdayThursday  -6.903e-02  8.816e-02  -0.783  0.43363    
created_at_weekdayTuesday   -2.766e-02  8.392e-02  -0.330  0.74167    
created_at_weekdayWednesday -1.607e-01  8.640e-02  -1.859  0.06297 .  
create2launch                2.185e-04  3.532e-04   0.619  0.53620    
launch2deadline             -1.348e-02  2.249e-03  -5.996 2.02e-09 ***
categoryApps                 1.645e+01  1.157e+03   0.014  0.98866    
categoryBlues                1.898e+01  1.157e+03   0.016  0.98692    
categoryComedy               3.262e+01  2.664e+03   0.012  0.99023    
categoryExperimental         1.752e+01  1.157e+03   0.015  0.98792    
categoryFestivals            1.732e+01  1.157e+03   0.015  0.98806    
categoryFlight               1.600e+01  1.157e+03   0.014  0.98897    
categoryGadgets              1.634e+01  1.157e+03   0.014  0.98873    
categoryHardware             1.650e+01  1.157e+03   0.014  0.98862    
categoryImmersive            1.749e+01  1.157e+03   0.015  0.98794    
categoryMakerspaces          1.665e+01  1.157e+03   0.014  0.98852    
categoryMusical              1.721e+01  1.157e+03   0.015  0.98813    
categoryPlaces              -3.240e-01  1.239e+03   0.000  0.99979    
categoryPlays                1.722e+01  1.157e+03   0.015  0.98813    
categoryRestaurants          6.951e-01  1.496e+03   0.000  0.99963    
categoryRobots               1.700e+01  1.157e+03   0.015  0.98828    
categoryShorts               3.244e+01  1.223e+03   0.027  0.97884    
categorySoftware             1.526e+01  1.157e+03   0.013  0.98948    
categorySound                1.697e+01  1.157e+03   0.015  0.98830    
categorySpaces               1.738e+01  1.157e+03   0.015  0.98802    
categoryThrillers            5.181e-02  1.498e+03   0.000  0.99997    
categoryUnknown              1.679e+01  1.157e+03   0.015  0.98842    
categoryWearables            1.656e+01  1.157e+03   0.014  0.98858    
categoryWeb                  1.488e+01  1.157e+03   0.013  0.98974    
```

```
categoryWebseries           -3.649e-01  1.430e+03   0.000  0.99980
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11556.1  on 8335  degrees of freedom
Residual deviance:  9556.7  on 8288  degrees of freedom
AIC: 9652.7

Number of Fisher Scoring iterations: 15
```

In [256…

```
# confidence interval
s = summary(lr)
coef_ci = data.frame(exp_coef = round(exp(s$coefficients[, 1]), 3),
                     lwr = round(exp(s$coefficients[, 1] - 1.96* s$coefficients
                     upr =  round(exp(s$coefficients[, 1] + 1.96* s$coefficient
                     p_value = round(s$coefficients[, 4], 3)
                    )
coef_ci
```

A data.frame: 48 × 4

	exp_coef	lwr	upr	p_value
	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	1.507844e+156	2.190865e+145	1.037761e+167	0.756
funding_goal1000	9.890000e-01	9.890000e-01	9.890000e-01	0.000
name_len	1.106000e+00	1.105000e+00	1.106000e+00	0.000
blurb_len	1.008000e+00	1.008000e+00	1.008000e+00	0.303
disable_communicationTrue	0.000000e+00	0.000000e+00	0.000000e+00	0.952
deadline_month	1.021000e+00	1.021000e+00	1.021000e+00	0.005
deadline_day	9.980000e-01	9.980000e-01	9.980000e-01	0.510
deadline_weekdayFriday	9.940000e-01	9.920000e-01	9.960000e-01	0.949
deadline_weekdaySaturday	9.070000e-01	9.050000e-01	9.090000e-01	0.330
deadline_weekdaySunday	8.130000e-01	8.110000e-01	8.150000e-01	0.037
deadline_weekdayThursday	1.090000e+00	1.088000e+00	1.092000e+00	0.373
deadline_weekdayTuesday	1.060000e+00	1.058000e+00	1.063000e+00	0.596
deadline_weekdayWednesday	9.990000e-01	9.970000e-01	1.001000e+00	0.993
created_at_year	8.300000e-01	8.290000e-01	8.300000e-01	0.000
created_at_month	9.840000e-01	9.840000e-01	9.840000e-01	0.041
created_at_day	9.980000e-01	9.980000e-01	9.980000e-01	0.455
created_at_weekdayFriday	8.710000e-01	8.690000e-01	8.730000e-01	0.132
created_at_weekdaySaturday	7.470000e-01	7.460000e-01	7.490000e-01	0.003
created_at_weekdaySunday	7.960000e-01	7.950000e-01	7.980000e-01	0.018
created_at_weekdayThursday	9.330000e-01	9.320000e-01	9.350000e-01	0.434
created_at_weekdayTuesday	9.730000e-01	9.710000e-01	9.740000e-01	0.742
created_at_weekdayWednesday	8.520000e-01	8.500000e-01	8.530000e-01	0.063
create2launch	1.000000e+00	1.000000e+00	1.000000e+00	0.536
launch2deadline	9.870000e-01	9.870000e-01	9.870000e-01	0.000
categoryApps	1.388129e+07	0.000000e+00	9.408436e+17	0.989
categoryBlues	1.744082e+08	3.000000e-03	1.182112e+19	0.987
categoryComedy	1.464667e+14	0.000000e+00	1.259001e+39	0.990
categoryExperimental	4.079178e+07	1.000000e-03	2.764778e+18	0.988
categoryFestivals	3.332464e+07	0.000000e+00	2.258671e+18	0.988
categoryFlight	8.870843e+06	0.000000e+00	6.012464e+17	0.989
categoryGadgets	1.247738e+07	0.000000e+00	8.456897e+17	0.989
categoryHardware	1.471637e+07	0.000000e+00	9.974431e+17	0.989
categoryImmersive	3.960699e+07	1.000000e-03	2.684476e+18	0.988

	exp_coef	lwr	upr	p_value
	<dbl>	<dbl>	<dbl>	<dbl>
categoryMakerspaces	1.703576e+07	0.000000e+00	1.154647e+18	0.989
categoryMusical	2.993899e+07	0.000000e+00	2.029199e+18	0.988
categoryPlaces	7.230000e-01	0.000000e+00	2.859881e+11	1.000
categoryPlays	2.998328e+07	0.000000e+00	2.032201e+18	0.988
categoryRestaurants	2.004000e+00	0.000000e+00	2.002286e+14	1.000
categoryRobots	2.412869e+07	0.000000e+00	1.635390e+18	0.988
categoryShorts	1.227256e+14	4.390510e+02	3.430481e+25	0.979
categorySoftware	4.258984e+06	0.000000e+00	2.886646e+17	0.989
categorySound	2.343724e+07	0.000000e+00	1.588525e+18	0.988
categorySpaces	3.518487e+07	1.000000e-03	2.384755e+18	0.988
categoryThrillers	1.053000e+00	0.000000e+00	1.101984e+14	1.000
categoryUnknown	1.964149e+07	0.000000e+00	1.331257e+18	0.988
categoryWearables	1.554714e+07	0.000000e+00	1.053751e+18	0.989
categoryWeb	2.908422e+06	0.000000e+00	1.971265e+17	0.990
categoryWebseries	6.940000e-01	0.000000e+00	1.676499e+13	1.000

In [257…
```
lr.probs =  predict(lr,type = "response")
```

In [258…
```
mean(ifelse(lr.probs > 0.5, 1, 0) == US.sampled$success)
```

```
0.691218809980806
```

In [259…
```
sum(lr.probs > 0.5)
```

```
4672
```

In [260…
```
index25000 = which(US.sampled$funding_goal > 25000 & lr.probs > 0.5)
head(US.sampled[index25000, ])
```

	project	funding_goal	name	name_len	blurb_len	pledged	backers	
	<int>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<int>	
208	1782322182	125000	Help Produce The Songs of Blind Willie Johnson	8	13	125154.24	956	suc
213	1402725713	27500	The Ice Queen	3	13	36229.00	353	suc
225	109131804	30000	Eimear Noone Presents "Songs of Zelda: A Link to the Celts"	11	17	46229.00	1243	suc
228	1318385098	50000	Rad Rodgers - The return of the 90's era Apogee platformer!	11	16	81861.97	3901	suc
233	1688269963	30000	Quantum Chess - #QuantumChess	4	16	32607.00	804	suc
234	846997546	75000	Saber Rider and the Star Sheriffs - 3DS / Steam / Dreamcast	12	13	96591.88	1072	suc

In [261…

```
length(index25000)
```

733

In [262…

```
summary(US.filtered[which(US.filtered$funding_goal > 25000), ])
```

```
   project           funding_goal        name              name_len     
Min.   :2.610e+05   Min.   : 25073   Length:4548        Min.   : 1.000  
1st Qu.:5.608e+08   1st Qu.: 40000   Class :character   1st Qu.: 4.000  
Median :1.095e+09   Median : 55000   Mode  :character   Median : 7.000  
Mean   :1.086e+09   Mean   : 90267                      Mean   : 6.226  
3rd Qu.:1.618e+09   3rd Qu.:100000                      3rd Qu.: 8.000  
Max.   :2.146e+09   Max.   :500000                      Max.   :15.000  
                                                                        
  blurb_len        pledged             backers               state     
Min.   : 1.00   Min.   :     0.0   Min.   :   0.0   canceled  : 774  
1st Qu.:11.00   1st Qu.:    54.8   1st Qu.:   2.0   failed    :2751  
Median :13.00   Median :  1724.5   Median :  17.0   live      : 125  
Mean   :12.94   Mean   : 29658.1   Mean   : 220.7   successful: 855  
3rd Qu.:15.00   3rd Qu.: 24600.5   3rd Qu.: 137.2   suspended :  43  
Max.   :30.00   Max.   :492204.0   Max.   :8776.0                    
                                                                     
   success      disable_communication deadline_year  deadline_month  
Min.   :0.000   False:4505            Min.   :2010   Min.   : 1.000  
1st Qu.:0.000   True :  43            1st Qu.:2014   1st Qu.: 4.000  
Median :0.000                         Median :2015   Median : 7.000  
Mean   :0.188                         Mean   :2015   Mean   : 6.744  
3rd Qu.:0.000                         3rd Qu.:2016   3rd Qu.:10.000  
Max.   :1.000                         Max.   :2017   Max.   :12.000  
                                                                     
 deadline_day     deadline_weekday created_at_year created_at_month
Min.   : 1.00   Friday   :903      Min.   :2010    Min.   : 1.000  
1st Qu.: 8.00   Monday   :472      1st Qu.:2014    1st Qu.: 4.000  
Median :15.00   Saturday :672      Median :2015    Median : 7.000  
Mean   :15.61   Sunday   :642      Mean   :2015    Mean   : 6.466  
3rd Qu.:23.00   Thursday :772      3rd Qu.:2015    3rd Qu.: 9.000  
Max.   :31.00   Tuesday  :400      Max.   :2017    Max.   :12.000  
                Wednesday:687                                      
created_at_day  created_at_weekday launched_at_year launched_at_month
Min.   : 1.00   Friday   :620      Min.   :2010     Min.   : 1.000   
1st Qu.: 8.00   Monday   :777      1st Qu.:2014     1st Qu.: 4.000   
Median :16.00   Saturday :425      Median :2015     Median : 7.000   
Mean   :15.73   Sunday   :434      Mean   :2015     Mean   : 6.558   
3rd Qu.:23.00   Thursday :709      3rd Qu.:2016     3rd Qu.:10.000   
Max.   :31.00   Tuesday  :833      Max.   :2017     Max.   :12.000   
                Wednesday:750                                        
launched_at_day launched_at_weekday create2launch     launch2deadline
Min.   : 1.00   Friday   : 589      Min.   :  0.00   Min.   : 1.00  
1st Qu.: 8.00   Monday   : 961      1st Qu.:  7.00   1st Qu.:30.00  
Median :15.00   Saturday : 182      Median : 24.00   Median :30.00  
Mean   :15.19   Sunday   : 184      Mean   : 54.82   Mean   :37.04  
3rd Qu.:23.00   Thursday : 665      3rd Qu.: 65.00   3rd Qu.:45.00  
Max.   :31.00   Tuesday  :1094      Max.   :495.00   Max.   :89.00  
                Wednesday: 873                                      
     category   
Hardware :1102  
Gadgets  : 744  
Web      : 596  
Software : 562  
Unknown  : 360  
Wearables: 348  
(Other)  : 836  
```

In [294…

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
```

```
}
tmp = c(date.vars, 'deadline_weekday', 'created_at_weekday', 'launched_at_weekc
for (v in tmp) {
    print(paste0(v, ': ', getmode(US.filtered[which(US.filtered$funding_goal >
}
```

```
[1] "deadline_year: 2015"
[1] "deadline_month: 12"
[1] "deadline_day: 1"
[1] "created_at_year: 2015"
[1] "created_at_month: 7"
[1] "created_at_day: 22"
[1] "launched_at_year: 2015"
[1] "launched_at_month: 10"
[1] "launched_at_day: 8"
[1] "deadline_weekday: Friday"
[1] "created_at_weekday: Tuesday"
[1] "launched_at_weekday: Tuesday"
```

In [295…

```
newx = data.frame(
    funding_goal1000 = c(25), ## -
    name_len = c(15), ## +
    blurb_len = c(13),
    disable_communication = c('False'),
    deadline_month = c(12), ## +
    deadline_day = c(1),
    deadline_weekday = c('Friday'), ## Sunday -
    created_at_year = c(2022), ## -
    created_at_month = c(1), ## -
    created_at_day = c(22),
    created_at_weekday = c('Tuesday'), ## Sunday and Sataurday -
    create2launch = c(24),
    launch2deadline = c(1), ## -
    category = c('Unknown')
)
```

In [296…

```
# predict
predict(lr, newdata = newx, type = 'response')
```

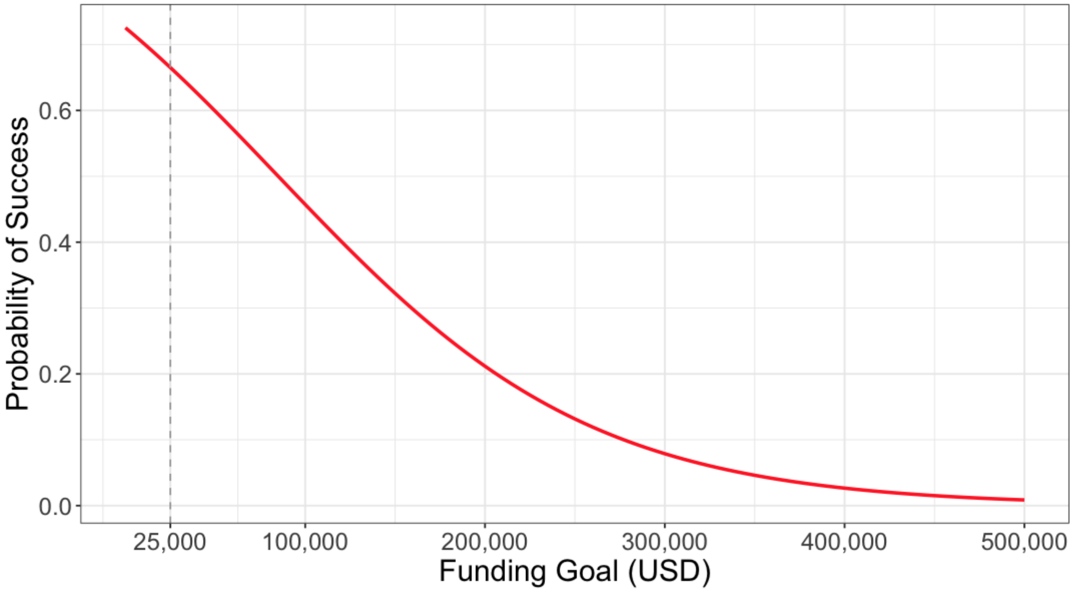
1: 0.665141485330049

In [297…

```
# visualize relationship between funding goals and probability of success
fg = seq(0, 500000, 5000)
prob = c()
for (x in fg) {
    newx$funding_goal1000 = x/1000
    prob = c(prob, predict(lr, newx, type = 'response'))
}
```

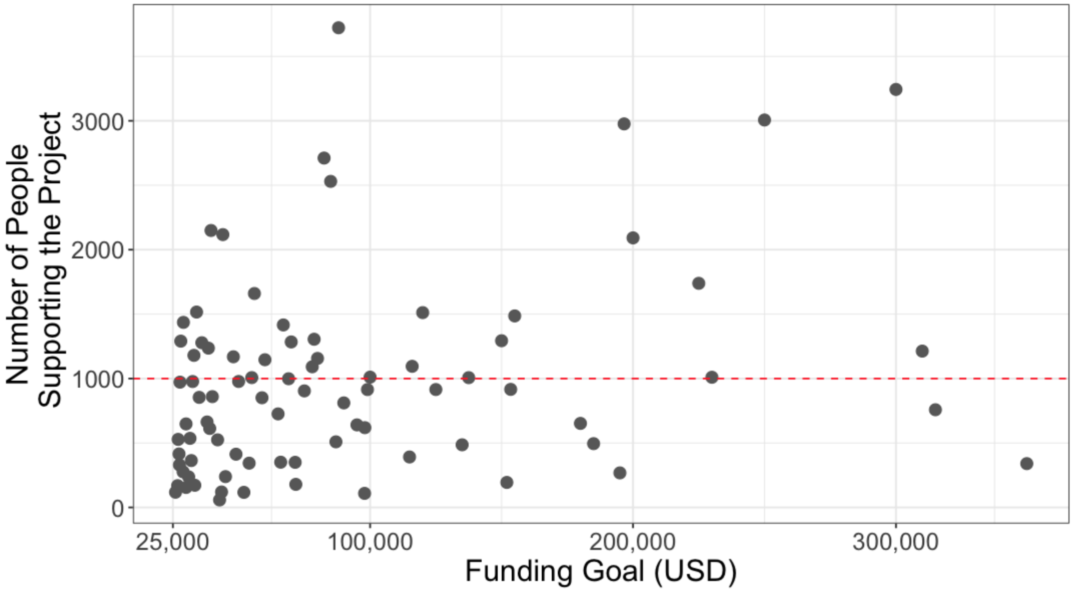
In [298…

```
options(repr.plot.width = 9, repr.plot.height = 5)
ggplot() +
    geom_line(aes(x = fg, y=prob), size = 1, color = 'firebrick1') +
    theme_bw() +
    theme(text = element_text(size = 18), legend.position = c(0.85, 0.85)) +
    scale_x_continuous(breaks = c(2.5e+4, 1e+5, 2e+5, 3e+5, 4e+5, 5e+5),
                       labels = c('25,000', '100,000', '200,000', '300,000', '4
    geom_vline(xintercept = 25000, linetype='dashed', color = 'darkgray') +
    xlab('Funding Goal (USD)') + ylab('Probability of Success')
```



In [277…

```
# visualize the number of backers
back.df = US.filtered[which(US.filtered$funding_goal > 25000 & US.filtered$succ
back.df = back.df %>% group_by(funding_goal) %>% summarise(backers.avg = mean(b
back.df %>% ggplot() +
    geom_point(aes(x = funding_goal, y=backers.avg), size = 3, color = 'dimgray
    theme_bw() +
    theme(text = element_text(size = 18), legend.position = c(0.85, 0.85)) +
    scale_x_continuous(breaks = c(2.5e+4, 1e+5, 2e+5, 3e+5, 4e+5, 5e+5),
                       labels = c('25,000', '100,000', '200,000', '300,000', '4
    geom_hline(yintercept = 1000, linetype='dashed', color = 'firebrick1') +
    xlab('Funding Goal (USD)') + ylab('Number of People\n Supporting the Projec
```



Logistic Regression for number of backers

In [278…

```
US25000 = US.filtered[which(US.filtered$funding_goal > 25000 & US.filtered$succ
US25000$backer1000 = as.integer(ifelse(US25000$backers >= 1000, 1, 0))
```

In [279…

```
head(US25000)
```

	project	funding_goal	name	name_len	blurb_len	pledged	backers	
	<int>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<int>	
208	1782322182	125000	Help Produce The Songs of Blind Willie Johnson	8	13	125154.24	956	su
213	1402725713	27500	The Ice Queen	3	13	36229.00	353	su
224	2143543297	150000	SPORTSFRIENDS featuring Johann Sebastian Joust	5	12	152451.25	4146	su
225	109131804	30000	Eimear Noone Presents "Songs of Zelda: A Link to the Celts"	11	17	46229.00	1243	su
228	1318385098	50000	Rad Rodgers - The return of the 90's era Apogee platformer!	11	16	81861.97	3901	su
229	1692978427	300000	OVERLOAD - The Ultimate Six-Degree-of-Freedom Shooter	6	7	306537.00	4896	su

In [300…

```
summary(US25000$backers)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   20.0   303.0   594.0   948.6  1216.0  8776.0 
```

In [280…

```
US25000$deadline_weekday = relevel(US25000$deadline_weekday, ref = 'Monday')
US25000$created_at_weekday = relevel(US25000$created_at_weekday, ref = 'Monday'
```

In [281…

```
lr2 = glm(backer1000 ~ funding_goal + name_len + blurb_len + #disable_communica
          deadline_month + deadline_day + deadline_weekday + 
          created_at_year + created_at_month + created_at_day + created_at_weekd
          create2launch + launch2deadline + category,
          data = US25000,
          family = "binomial")
```

In [282…

```
summary(lr2)
```

```
Call:
glm(formula = backer1000 ~ funding_goal + name_len + blurb_len + 
    deadline_month + deadline_day + deadline_weekday + created_at_year + 
    created_at_month + created_at_day + created_at_weekday + 
    create2launch + launch2deadline + category, family = "binomial", 
    data = US25000)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8498  -0.8846  -0.6449   1.1525   2.2290  

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 3.106e+02  1.386e+02   2.242   0.0250 *  
funding_goal                9.365e-06  1.857e-06   5.044 4.56e-07 ***
name_len                    1.742e-02  3.485e-02   0.500   0.6171    
blurb_len                  -2.154e-02  2.722e-02  -0.791   0.4287    
deadline_month              2.967e-02  2.421e-02   1.225   0.2205    
deadline_day                1.208e-02  9.024e-03   1.338   0.1808    
deadline_weekdayFriday      4.902e-01  3.004e-01   1.632   0.1027    
deadline_weekdaySaturday    1.397e-01  3.299e-01   0.424   0.6719    
deadline_weekdaySunday      2.902e-01  3.376e-01   0.860   0.3899    
deadline_weekdayThursday    1.913e-01  3.036e-01   0.630   0.5286    
deadline_weekdayTuesday    -4.907e-01  3.963e-01  -1.238   0.2157    
deadline_weekdayWednesday  -3.381e-01  3.369e-01  -1.004   0.3155    
created_at_year            -1.550e-01  6.877e-02  -2.254   0.0242 *  
created_at_month           -1.496e-02  2.554e-02  -0.586   0.5580    
created_at_day              3.021e-03  9.539e-03   0.317   0.7515    
created_at_weekdayFriday   -3.611e-01  2.951e-01  -1.224   0.2211    
created_at_weekdaySaturday  1.561e-01  3.558e-01   0.439   0.6608    
created_at_weekdaySunday    3.691e-01  3.422e-01   1.079   0.2808    
created_at_weekdayThursday -2.937e-01  2.826e-01  -1.039   0.2986    
created_at_weekdayTuesday  -5.553e-03  2.627e-01  -0.021   0.9831    
created_at_weekdayWednesday 4.288e-01  2.748e-01   1.560   0.1187    
create2launch              -1.536e-03  1.048e-03  -1.466   0.1427    
launch2deadline            -5.862e-03  9.551e-03  -0.614   0.5394    
categoryBlues              -1.577e+01  1.652e+03  -0.010   0.9924    
categoryExperimental        1.762e+01  2.400e+03   0.007   0.9941    
categoryFestivals          -1.613e+01  2.400e+03  -0.007   0.9946    
categoryFlight             -6.938e-01  7.567e-01  -0.917   0.3592    
categoryGadgets             6.489e-01  5.037e-01   1.288   0.1976    
categoryHardware           -2.333e-02  4.970e-01  -0.047   0.9626    
categoryImmersive          -1.527e+01  8.169e+02  -0.019   0.9851    
categoryMakerspaces        -1.558e+01  9.394e+02  -0.017   0.9868    
categoryMusical             7.521e-01  7.952e-01   0.946   0.3443    
categoryPlays              -1.514e+01  7.338e+02  -0.021   0.9835    
categoryRobots             -2.190e-01  6.083e-01  -0.360   0.7188    
categoryShorts             -1.597e+01  2.400e+03  -0.007   0.9947    
categorySoftware            2.147e-01  6.663e-01   0.322   0.7473    
categorySound              -3.530e-01  6.048e-01  -0.584   0.5595    
categorySpaces              5.699e-01  8.024e-01   0.710   0.4775    
categoryUnknown             1.089e+00  5.165e-01   2.108   0.0350 *  
categoryWearables           7.303e-01  5.169e-01   1.413   0.1577    
categoryWeb                -5.070e-01  8.020e-01  -0.632   0.5273    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1075.53  on 854  degrees of freedom
```

```
Residual deviance:  958.75  on 814  degrees of freedom
AIC: 1040.7

Number of Fisher Scoring iterations: 15
```

In [283…

```
# predict the probability
newx = data.frame(
    funding_goal = c(25000), ## -
    name_len = c(15), ## +
    blurb_len = c(13),
    disable_communication = c('False'),
    deadline_month = c(12), ## +
    deadline_day = c(16),
    deadline_weekday = c('Friday'), ## Sunday -
    created_at_year = c(2022), ## -
    created_at_month = c(1), ## -
    created_at_day = c(16),
    created_at_weekday = c('Wednesday'), ## Sunday and Sataurday -
    create2launch = c(55),
    launch2deadline = c(1), ## -
    category = c('Unknown')
)
```

In [284…

```
predict(lr2, newx, 'response')
```

1: 0.463227788910996

In [285…

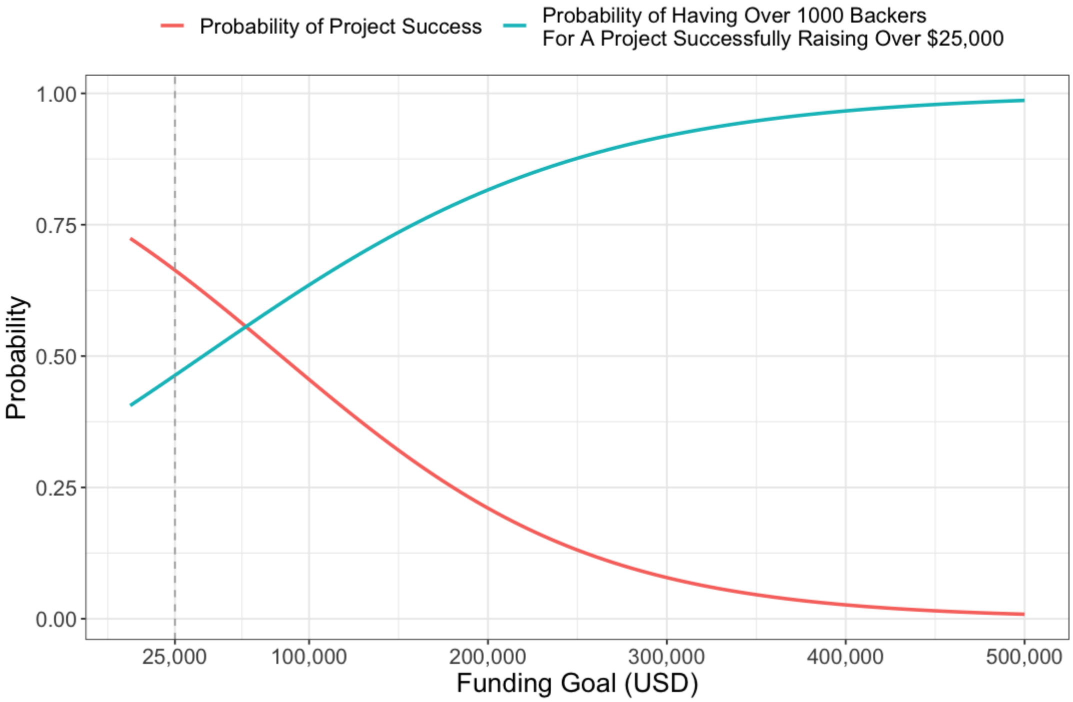
```
# visualize relationship between funding goals and probability of success
fg = seq(0, 500000, 5000)
probb = c()
for (x in fg) {
    newx$funding_goal = x
    probb = c(probb, predict(lr2, newx, type = 'response'))
}
```

In [286…

```
plt.df = data.frame(fg = fg, prob = prob, probb = probb)
plt.df = plt.df %>% gather(key = 'type', value = 'prob', -c('fg'))
```

In [289…

```
options(repr.plot.width = 9, repr.plot.height = 6)
plt.df %>% ggplot() +
    geom_line(aes(x=fg, y=prob, color=type), size = 1) +
    theme_bw() +
    theme(text = element_text(size = 16), legend.position = "top", legend.box =
    scale_x_continuous(breaks = c(2.5e+4, 1e+5, 2e+5, 3e+5, 4e+5, 5e+5),
                       labels = c('25,000', '100,000', '200,000', '300,000', '4
    geom_vline(xintercept = 25000, linetype='dashed', color = 'darkgray') +
    scale_color_discrete(name = "",
                         labels = c('Probability of Project Success',
                                    'Probability of Having Over 1000 Backers\nE
    xlab('Funding Goal (USD)') + ylab('Probability')
```


Probability of Project Success
Probability of Having Over 1000 Backers
For A Project Successfully Raising Over $25,000
Probability
1.00
0.75
0.50
0.25
0.00
25,000
100,000
200,000
300,000
400,000
500,000
Funding Goal (USD)

```
In [ ]:
```