# STATISTICS DISTRIBUTIONS

The summary of variable statistics is shown below:

**Numeric variables**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Rating | Price | Team size | Coin Num | Distributed Percentage |
| Min | 1.00 | 0.00 | 1.00 | 1.200e+01 | 0.00 |
| 1 Qua | 2.60 | 0.04 | 7.00 | 5.000e+07 | 0.40 |
| Median | 3.10 | 0.12 | 12.00 | 1.800e+08 | 0.55 |
| Mean | 3.12 | 19.01 | 13.11 | 8.178e+12 | 1.06 |
| 3rd Qu | 3.70 | 0.50 | 17.00 | 6.000e+08 | 0.70 |
| Max | 4.80 | 39384.00 | 75.00 | 2.262e+16 | 869.75 |
| NA’s (null values) | 0.00 | 180.00 | 154 | 0.00 | 0.00 |

Understanding the different variables, it is evident that certain variables like price and distributed percentage exhibit outliers, while price and team size have missing values. Coin number has larger quantities which also show outliers.

It is also worth noting that distributed percentage variable represents percentage values in decimals. The range is between 0 to 1. A maximum of 869.75 value is therefore incorrect. These will be handled in the later steps under data preparation.

**Categorical variables**

|  |  |
| --- | --- |
| Has Video *(dummy)* |  |
| 0 | 1 |
| 758 | 2009 *(72.60%)* |

|  |  |
| --- | --- |
| Has GitHub *(dummy)* |  |
| 0 | 1 |
| 1168 | 1599 *(57.78%)* |

|  |  |
| --- | --- |
| Has Reddit *(dummy)* |  |
| 0 | 1 |
| 1016 | 1751 *(63.62%)* |

|  |  |
| --- | --- |
| Min Investment *(dummy)* |  |
| 0 | 1 |
| 1513 | 1254 *(45.31%)* |

Overall, all features have a fair distribution of categories, expect *‘has video’*. It represents that about 72% of the project campaigns have videos describing their project. Over 57% of project campaigns have GitHub available, 63% has reddit, and only 45% projects require minimum investment threshold. Hence majority of projects have online presence and engagement, but majority of them do not have minimum investment criteria.

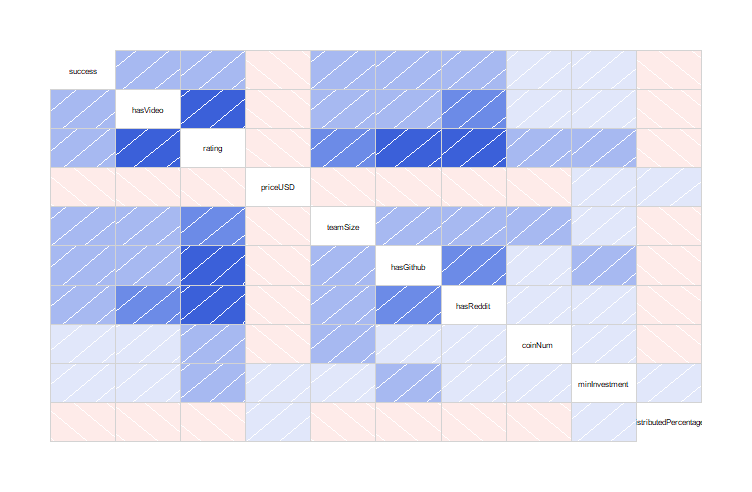
**Dependent variable:** -

|  |  |
| --- | --- |
| Success |  |
| “Y” = Yes | **“N” = No** |
| 1739 | **1028** |

“success” variable signifies if the company/ project has successful completed its ICO campaign or not. The variable has 62-37 % split between its classes. Hence, data imbalance may not be an issue while training the ML model.

# DATA VIZUALIZATION

Further exploring the relationship of variables with dependent variable:

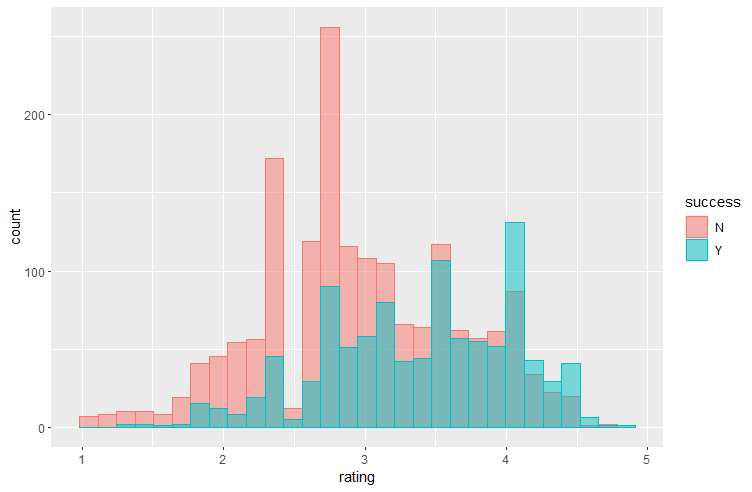
The corelation matrix shows a positive corelation between dependent variable ‘success’ and online engagement (‘hasVideo’, ‘hasGitHub’, ‘hasReddit’). It is also evident that having online presence has a very strong corelation with ratings.

The dependent variable also shows positive corelation with team size and a weak positive relation with coin numbers and minimum investment

Distributed percentage & price have a negative corelation with success outcome of the project. The more percentage share is for investors, the lesser probability of success. Having lesser share of the project within internal team could be seen less trustworthy and riskier to invest in such projects.

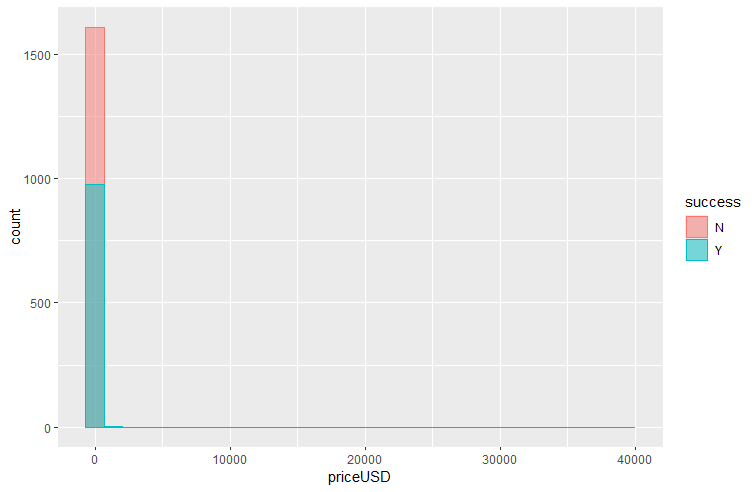
Visualising the distribution with other variables:

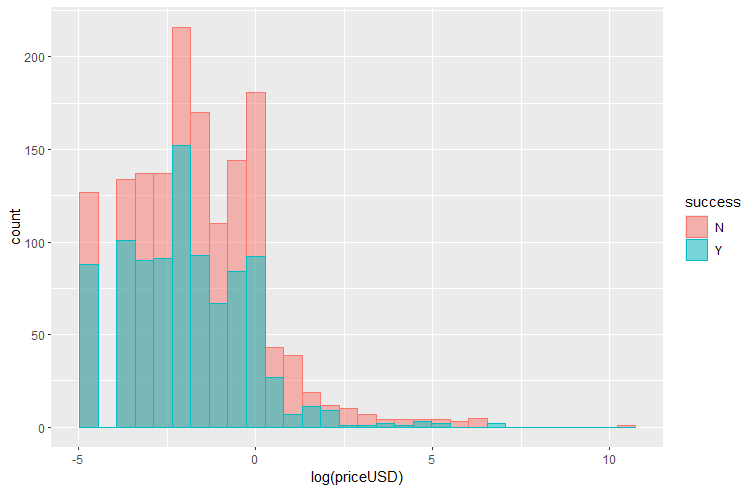
* Ratings



There is uniform distribution of success and not success with ratings variable.

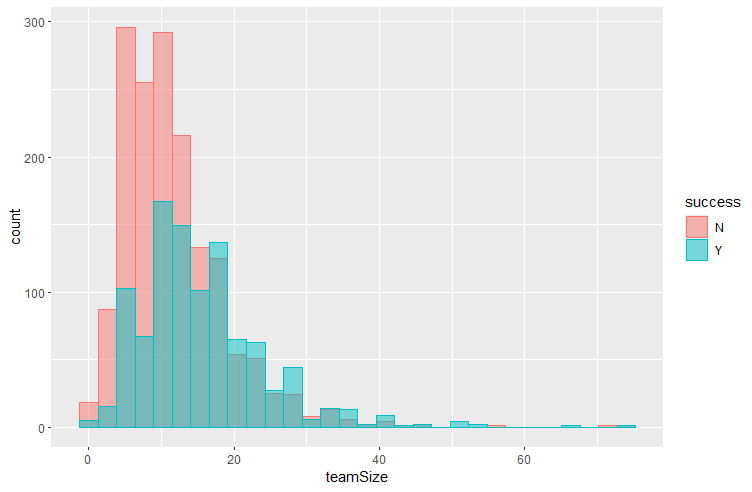
* Price (in USD)



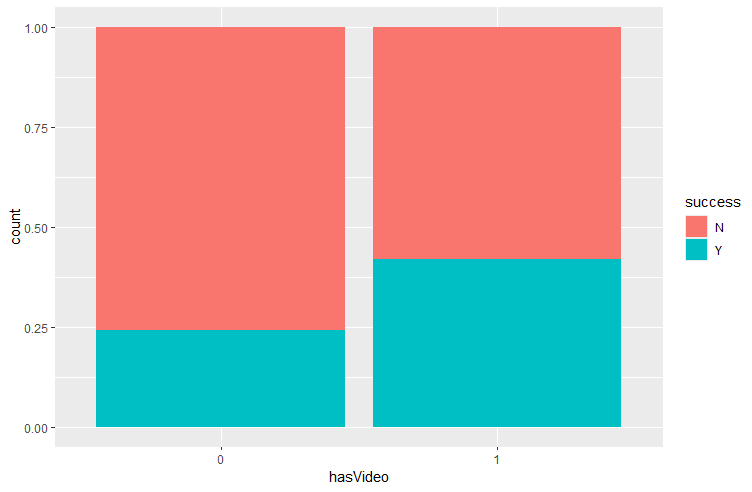


There exist many outliers in the data. Log transformation is applied to view the distribution.

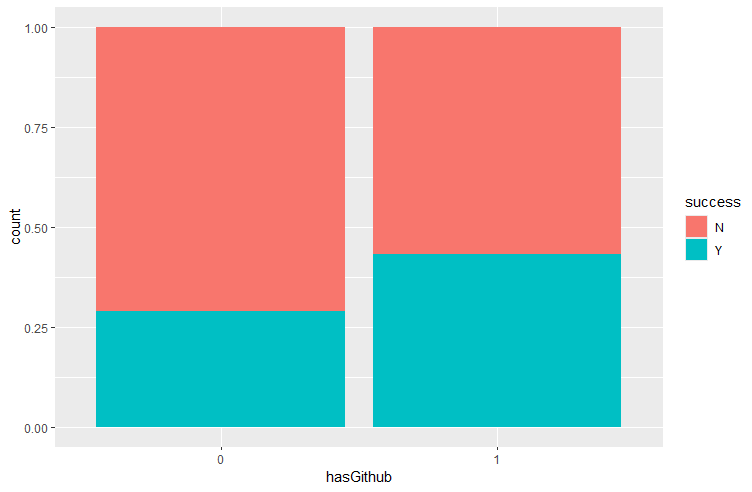
* Team Size



* Has Video



* Has GitHub



* Has Reddit



* Minimum Investment

