* Use your data to determine whether the mean or the median better summarizes the data.

The calculated statistical table for the number of backers for successful and unsuccessful(failed) projects have been given in Table 1. It can be concluded that we have an asymmetric dataset in both the cases successful as well as the failed ones with respect to the number of backers. Median would better summarize the dataset as the extreme data values (outliers) are contributing to the mean, which skewed the mean in a right direction if we compare with their respective median values.

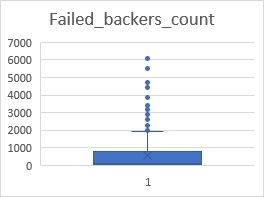
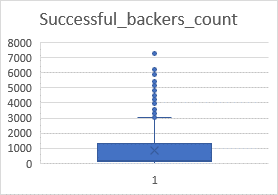
 

Table 1: Statistical analysis results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcome Successful** |  | **Outcome Failed** |  |
| Mean number of backers | 851 | Mean number of backers | 586 |
| Median number of backers | 201 | Median number of backers | 115 |
| Minimum number of backers | 16 | Minimum number of backers | 0 |
| Maximum number of backers | 7295 | Maximum number of backers | 6080 |
| Variance number of backers | 1603374 | Variance number of backers | 921575 |
| Standard deviation number of backers | 1266 | Standard deviation number of backers | 960 |
| First quartile | 128 | First quartile | 38 |
| Second quartile (median) | 201 | Second quartile (median) | 115 |
| Third quartile | 1289 | Third quartile | 790 |
| Interquartile range (IQR) | 1161 | Interquartile range (IQR) | 752 |
| 1.5 \* IQR | 1742 | 1.5 \* IQR | 1127 |
| Low end outliers | -1614 | Low end outliers | -1089 |
| High end outlier | 3030 | High end outlier | 1917 |

* Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?

The variance is low for failed number of backers as compared to successful ones. The higher the variance the more distributed the data is. But in this case where the dataset has been skewed in a particular direction (right skewed) via significant outlier presence, Variance or standard deviation won’t be the best to talk about the spread or distribution of the data points. We have already concluded that Median is a better way to summarize the data in these cases, so Interquartile range (IQR) will be the robust way to tell the data distribution. The IQR is not going to be affected by the presence of outliers.

For successful number of backers as shown in Table 1:

Min = 16, IQR1 = 128, IQR2 = 201, IQR3 = 1289, Max = 7295

Outliers = [-1614, 3030]

Any data point of backers\_count which is lower than -1614 and higher than 3030 are outliers. We see from the data or box graph that there is a lot of data point which are above 3030. We see significant outliers in this case.

For unsuccessful number of backers as shown in Table 1:

Min = 0, IQR1 = 38, IQR2 = 115, IQR3 = 789, Max = 6080

Outliers = [-1089, 1917]

Similarly, any data point of backers\_count which is lower than -1089 and higher than 1917 are outliers. We see from the data or box graph that there is a lot of data point which are above 1917, making the outliers.