**CHALLENGE 1**

**REPORT**

**(Answers in Blue)**

\*\*\*

* Create a report in Microsoft Word, and answer the following questions:
  + Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?

The data analyzed and visualized can be useful in determining trends associated with crowdfunding campaigns.

* + 1. From the chart Categories vs. Outcomes, the most successful campaigns were in **audio-visual entertainment** categories. Surprisingly, theater in particular was the subject of the most successful crowdfunding campaigns.

Failures were also high in the same categories, although not as high as successes. Numbers of cancelled and live campaigns are relatively much lower than both successful and failed campaigns.

Crowdfunding projects in sub-categories demonstrate the same trend in outcomes. Sub-categories with high successes also account for high failures. Audio-visual entertainment industry dominates the field of successful outcomes.

* + 1. From the chart for Outcomes based on financial goals, it can be observed that crowdfunding projects with goals between $15,000 and $50,000 had **~ 85% success rate on average**, with 100% success rate for projects funded between $15,000 and $25,000.

This implies that failure rates would be low in these intervals, as is seen in the chart. Since canceled projects make up under 3% of all crowdfunded projects, successful projects and failed projects show **complementary trends** with respect to financial goals of the projects.

* + 1. From the chart “Launch-month vs. Outcomes”, it can be seen that successful campaigns were between 40% - 50% for any launch-time throughout any year between 2010 and 2020. However, summer months of **May, June and July** indicate an increasing trend, perhaps due to many countries being on leisure during these months and preference for entertainment increases.
  + What are some limitations of this dataset?
* The dataset includes crowdfunding projects in several countries, though mostly in the USA. The currencies do not have equivalent values; therefore, comparisons cannot be made reliably.
* It is difficult to make independent comparisons of success rate with the total contribution and success rate and number of backers, as they both influence the success of a campaign. Success rate should be a **combination** of funding and well as number of backers, regardless of the “Goal”.
  + What are some other possible tables and/or graphs that we could create, and what additional value would they provide?
* **Box plots** of “success” and “failure” (plotted here for the statistics) show the influence of extreme outliers on some statistical parameters, such as mean and standard deviation.
* **Average donation per backer vs. outcomes** (successful, failed, cancelled) would provide some insight into the financial ability of participants and their interests.
* Additional plot of categories of **outcomes by year from 2010 to 2020**, to assess the cumulative outcome per year to understand the effectiveness of crowdfunding over a period of 10 years.
* Use your data to determine whether the mean or the median better summarizes the data.

**Boxplots**

A picture containing text, screenshot, number, line

Description automatically generated

From the statistical parameters calculated for “**successful**” outcomes, mean is 851 and median is 201. Because of the presence of a large number of outliers, each outlier representing a high number of “**backers**”, the distribution is skewed to the right.

In this case, the **mean** is influenced by the outliers and is shifted to crowdfunding campaigns with larger numbers of “**backers**”.

Median, being the value of backers at 50% of the projects, it is less affected by outliers.

Therefore, **median** could be considered as the **central tendency** to describe this data set.

Similarly, for “**failed**” outcomes, mean is 586 and median is 115. Here too, outliers, each with number of “**backers**” shift the value of mean toward a larger value compared to the median. Thus, while mean is affected by outliers, median is less affected by outliers. Hence, the choice of **median** as a stable central parameter representing the data.

<https://medium.com/geekculture/what-are-right-skewed-and-left-skewed-distributions-a29b3def7598>

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Success\_Count** | **Failure\_Count** | **Normalized Failure** |
|  | Project Count = | 565 | 364 | 560 |
|  | Mean = | 851 | 586 | 902 |
|  | Variance = | 1603374 | 921575 | 1419225 |
|  | Stdev = | 1266 | 960 | 1478 |
|  | Mode= | 85 | 1 | 1.54 |
|  | Min (Q0) = | 16 | 0 | 0 |
|  | Q1 = | 128 | 38 | 58 |
|  | Median (Q2) = | 201 | 115 | 177 |
|  | Q3 = | 1280 | 785 | 1209 |
|  | Max (Q4) = | 7295 | 6080 | 9363 |
|  | IQR = | **1152** | **747** | **1150** |
|  |  |  |  |  |
|  | **Outliers** | 42 | 32 | 49 |

Standard deviation (σ) is a measure of dispersion of data. Observing the data above, σ for “success” (1260) is larger than that of σ for “failed” (960). This may give the impression that “success” data shows more variability compared to “failed” data.

There are more successful campaigns than failed projects, since there are more backers per successful project. But there are also successful projects with smaller numbers of backers, as there are failed campaigns with large number of backers.

Also, both sets of data show different ranges.

Therefore, for a fair comparison, both sets of data should be on the same scale. **Normalization** is one of the methods of standardization.

If the normalizing criterion were IQR (Inter-Quartile Range), values in “failed” would be multiplied by 1.54.

Now the “success” data and the “normalized failure” data become comparable.

σ for the “normalized failure” data becomes larger than that of the “success” data.

Now it can be argued that effectively the “failed” data shows more variability compared to the “success” data.