Domain: Startup Ecosystem

Overview:

Company X, a prominent Indian digital publisher focused on providing deep insights into the startup landscape, is committed to equipping its readers with practical and valuable knowledge. In the fast-paced and ever-evolving world of startups, the company understands the importance of identifying key financial factors that distinguish successful, ongoing startups from those that have shut down.

Objective:

This project aims to explore and determine whether there is a statistically significant difference in the average amount of funds raised by startups that are still in operation compared to those that have closed. Additionally, it seeks to examine whether a significant difference exists in the number of funding rounds between startups that continue to operate and those that have ceased operations.

Note:

*- Dataset Credits -->
https://www.kaggle.com/datasets/yanmaksi/big-startup-secsees-fail-dataset-from-crunchbase

Data Cleaning and Preparation

After loading the dataset, this is how it looked:
The main columns for the Statistics were funding_rounds, funding_total_usd and status

```
startup_df = pd.read_csv("D:/big_startup_secsees_dataset.csv")
```

```
startup_df.head()
```

	permalink	name	homepage_url	category_list	funding_total_usd	status	country_code	state_code	region	city	funding_rounds
0	/organization/-fame	#fame	http://livfame.com	Media	10000000	operating	IND	16	Mumbai	Mumbai	1
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms| Real Time| Social Network...	700000	operating	USA	DE	DE - Other	Delaware City	2
2	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games| Mobile	3406878	operating	NaN	NaN	NaN	NaN	1
3	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	2000000	operating	CHN	22	Beijing	Beijing	1

Statistics and Startups

As I was analyzing data for **indian** startups, I put a filter on the **country**, and for simplicity, I only wanted two categories under the status column which were **operating** and **closed**.

```
filtered_df = startup_df[
    (startup_df['country_code'] == 'IND') &
    (startup_df['status'].isin(['operating', 'closed']))
]
```

We had around 1467 Operating Startups and 69 Closed ones.

```
status_counts = filtered_df['status'].value_counts()
print(status_counts)
```

```
status
operating    1467
closed         69
Name: count, dtype: int64
```

Some info about the cities in which these startups were:

```
city_counts = filtered_df['city'].value_counts()
```

```
city_counts
```

```
city
Bangalore    372
Mumbai       282
New Delhi    170
Gurgaon      133
Hyderabad     84
```

Statistics and Startups

We had startups from multiple categories here:
Some categories were **E-Commerce, Software, Education, Internet, Finance, Real Estate** etc

```
num_categories = filtered_df['category_list'].nunique()
print("Number of unique categories:", num_categories)
```

```
Number of unique categories: 806
```

```
top_categories = filtered_df['category_list'].value_counts().head(10).index.tolist()


df_top_categories = filtered_df[filtered_df['category_list'].isin(top_categories)]


startup_counts_by_category = df_top_categories['category_list'].value_counts()

print(startup_counts_by_category)
```

```
category_list
E-Commerce          71
Software            67
Education           45
Internet            28
Finance             27
Real Estate         25
Mobile              24
Clean Technology    22
Curated Web         22
Apps                20
Name: count, dtype: int64
```

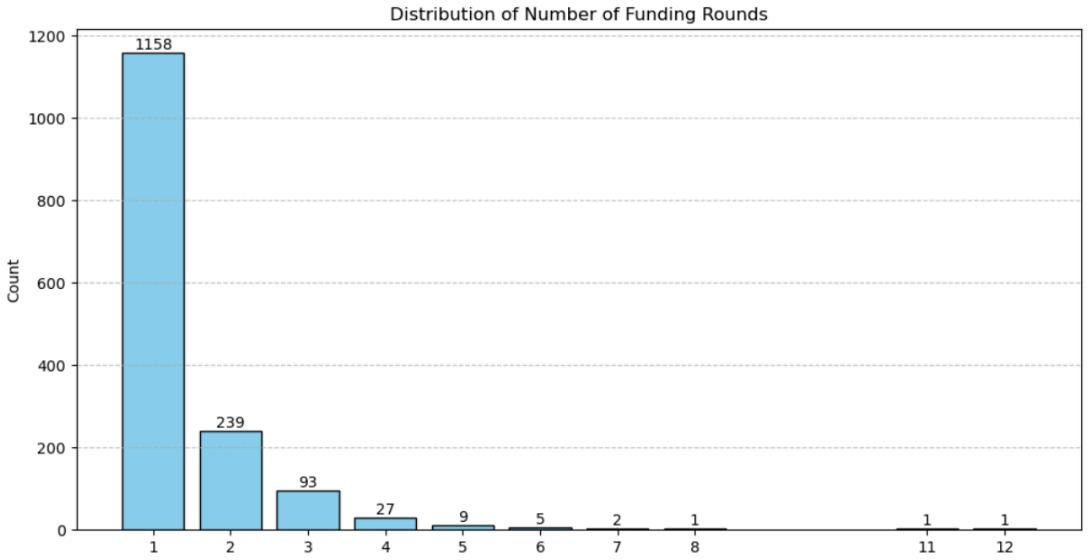
The below image reveals that the **majority** of the startups just went through **1** funding round.

```
funding_round_counts
```

```
funding_rounds
1     1158
2      239
3       93
4       27
5        9
6        5
7        2
8        1
11       1
12       1
Name: count, dtype: int64
```

Statistics and Startups

The Visualization of data based on **funding rounds** reveals that **1158 Startups** went through just **one** funding round.



Had to do some data cleaning here because the funding_total_usd (the amount of funding received) had some missing values and Nan data.

So we can see here that **795** startups went through **1** funding round.

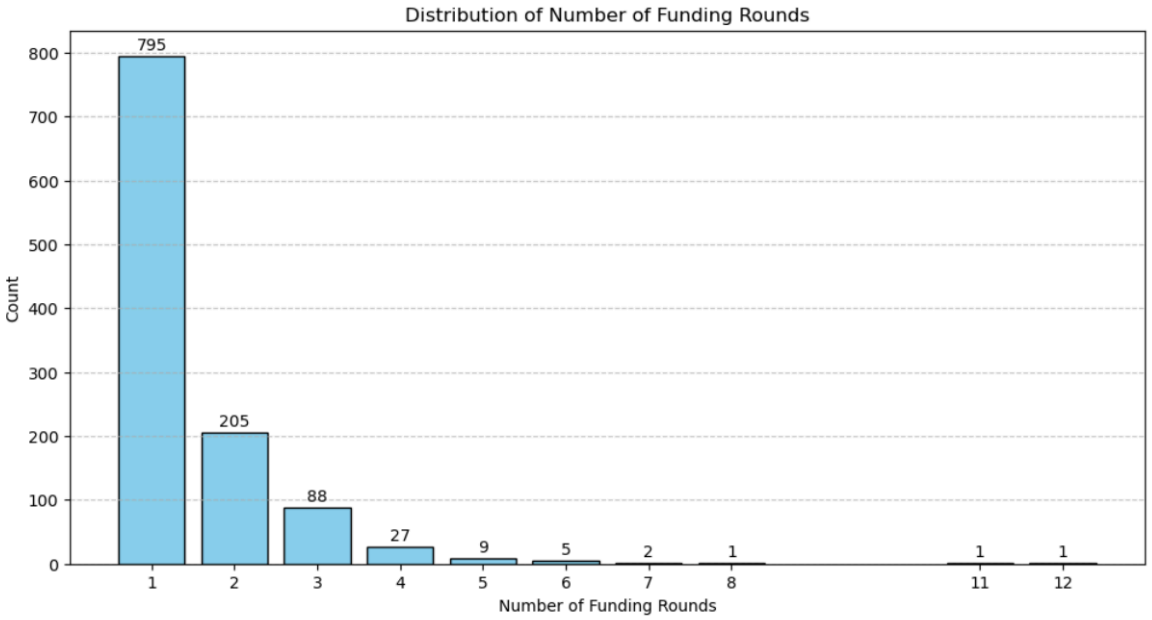
```
]: df_cleaned = filtered_df.dropna(subset= ['funding_total_usd
```

```
]: df_cleaned['funding_rounds'].value_counts()
```

```
]: funding_rounds
1     795
2     205
3      88
4      27
5       9
6       5
7       2
12      1
11      1
8       1
Name: count, dtype: int64
```

Statistics and Startups

So we can see here the visualization of data about the funding rounds which reveals that **795** startups went through **1** funding round.



Some details about the **amount of funding** received by startups:

```
filtered_df_1['funding_total_usd'].describe().map('${:,.2f}'.format)
```

```
count            $1,134.00
mean        $23,391,192.89
std        $153,640,841.53
min                $569.00
25%            $200,000.00
50%          $1,275,000.00
75%         $10,000,000.00
max      $3,151,140,000.00
Name: funding_total_usd, dtype: object
```

I then tried to find out the startups which had **Min** and **Max** amounts of funding:

So one was **Rural Server** which just got **$569** and the **Max** was **Flipkart** with **$3,151,140,000** which is a massive E-commerce Giant in India, which has its majority stake owned by Walmart.

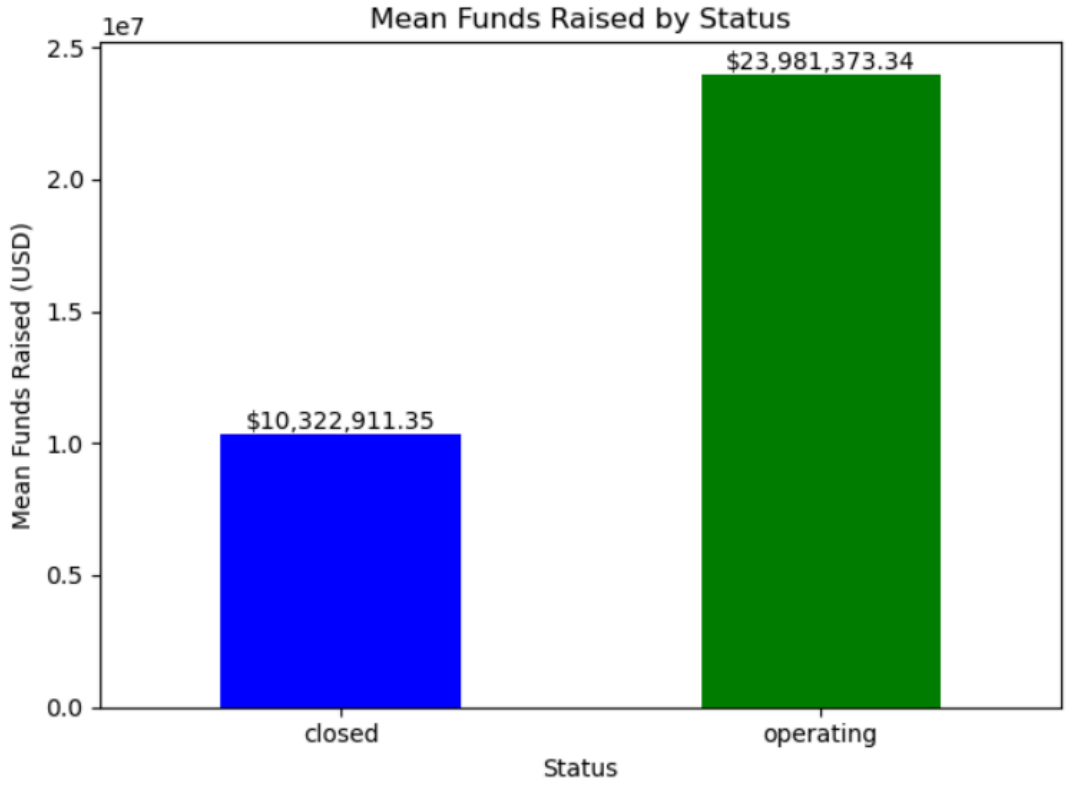
```
Startup with Minimum Funding: 569.0
              name  funding_rounds    city
49091  RuralServer               1   Noida

Startup with Maximum Funding: 3151140000.0
           name  funding_rounds       city
20874  Flipkart              12  Bangalore
```

Statistics and Startups

The code for finding out the max and min of the amount of funding data

```
min_value = filtered_df_1['funding_total_usd'].min()
max_value = filtered_df_1['funding_total_usd'].max()
mi = filtered_df_1[filtered_df_1['funding_total_usd'] == min_value]
mx = filtered_df_1[filtered_df_1['funding_total_usd'] == max_value]



# Display specific columns for the startup with minimum funding
min_funding_info = mi[['name', 'funding_rounds', 'city']]
print("Startup with Minimum Funding:", min_value)
print(min_funding_info)

# Display specific columns for the startup with maximum funding
max_funding_info = mx[['name', 'funding_rounds', 'city']]
print("\nStartup with Maximum Funding:", max_value)
print(max_funding_info)
```

Visualization of the Mean funds raised by startups that had closed down and were still operating



Statistics and Startups

The Statistics Part of the Report:

Mean Funds Raised

Null Hypothesis (H0): There is no statistically significant difference in the mean funds raised by currently operating startups and startups that have closed.

Alternative Hypothesis (H1): There is a statistically significant difference in the mean funds raised by currently operating startups and startups that have closed.

Using an independent sample t-test in this context is justified for the following reasons:

1. Comparing Two Independent Groups
2. Continuous Numeric Data
3. Normal Distribution Assumption
4. Homogeneity of Variance: We must check that using the **Levene Test**.

The Levene's Test is a statistical test used to assess whether the variances of two or more groups are equal or homogenous.

Null Hypothesis (H0): The null hypothesis in Levene's Test is that there are no significant differences in the variances of the groups being compared. In other words, it assumes that the variances are equal across all groups.

Alternative Hypothesis (H1): The alternative hypothesis in Levene's Test is that there are significant differences in the variances of the groups being compared. If the p-value is sufficiently small, you would reject the null hypothesis in favour of the alternative, indicating that at least one group has a significantly different variance compared to the others.

```
Levene's Test Statistic: 0.36074537025282777
P-value: 0.5482127964683872
Fail to reject the null hypothese: Vairances are equal.
```

The above results have determined that the variances are equal, thus meeting the assumption for the independent sample t-test. We then proceeded with the independent sample t-test.

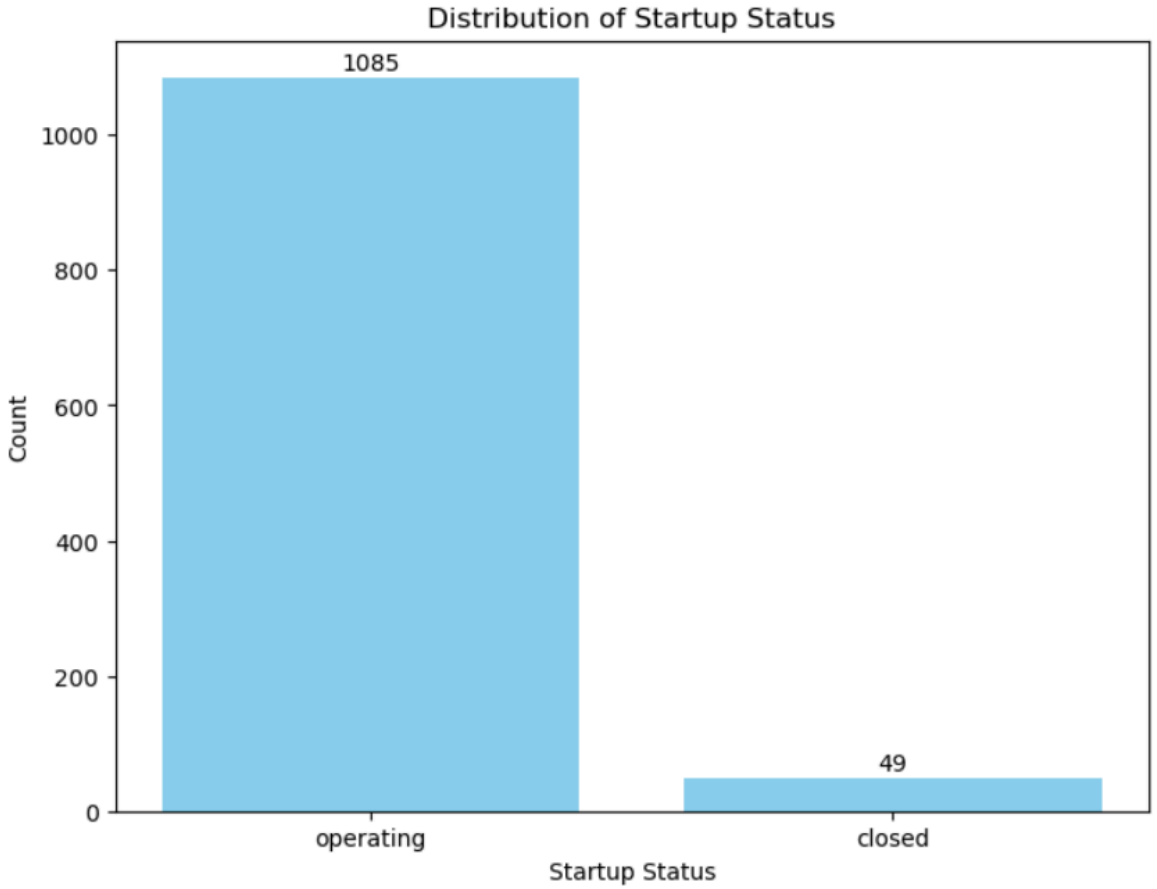
```
t, pvalue = ttest_ind(df3['Funds_Sample1'], df3['Funds_Sample2'], nan_policy = 'omit')
print('T-statistic',t)
print('P-value:', pvalue)
```

```
T-statistic 0.6085283061630911
P-value: 0.5429592211146083
```

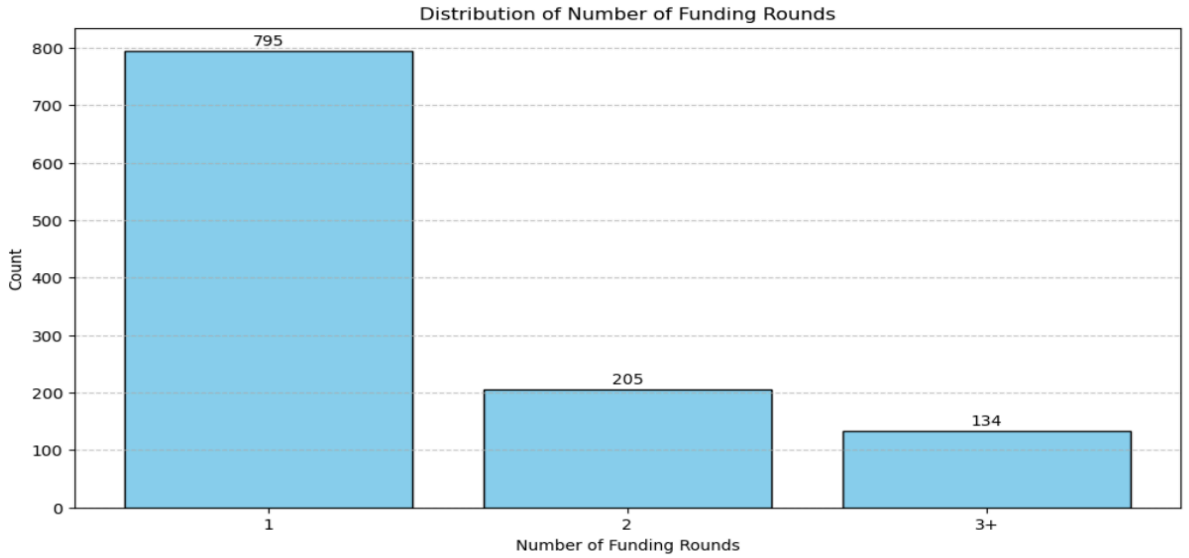
The p-value, which is greater than the chosen alpha level (i.e., 0.54 > 0.05), leads us to fail to reject the null hypothesis with 95% confidence. Therefore, we conclude that there is no statistically significant difference in the funds raised between currently operating startups and closed startups.

Statistics and Startups

Visualization based on the **status** of startups which we are analyzing, **1085** were **operating** and **49** had **closed** down their **operations.**



To facilitate analysis and meet the assumptions required for the **chi-square test**, we categorized the rounds of funding into three groups: **1, 2, and 3+**. This simplification helps ensure that the expected frequencies in each category meet the necessary **threshold of 5**, which is a key requirement for the validity of the chi-square test. These categories are reflected in the visualization below.



Statistics and Startups

All the values in the **expected frequencies exceed the threshold of 5**, indicating that our dataset meets the assumption of expected cell frequencies for the **Chi-Square Test of Independence**. So I proceeded with the **Hypothesis testing**.

```
Expected Frequencies:
status                          closed   operating
rounds of funding category
1                           34.351852  760.648148
2                            8.858025  196.141975
3+                           5.790123  128.209877
```

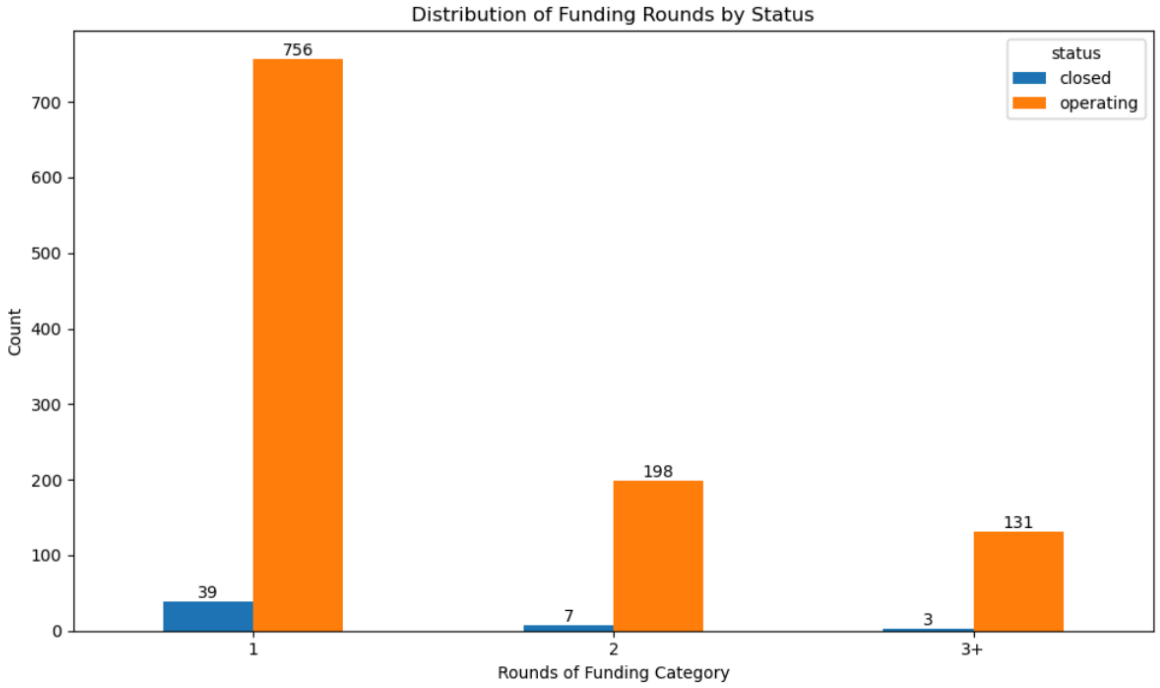
```
from scipy.stats import chi2_contingency

# Run the Chi Square Test

chi2, pval, dof, exp_freq = chi2_contingency(contingency_table, correction = False)
print('The p-value is',pval)
```

```
The p-value is 0.2908506204110049
```

As the **P-value is 0.29> 0.05** as a result we **fail** to reject the **null hypothesis**. This suggests that there is no statistically significant difference in the **number of funding rounds** between currently **operating** startups and startups that have **closed** their operations based on our chosen level of significance which was 0.05.

Statistics and Startups

A Visualization based on the number of startups which were either operating or closed down their operations and the rounds of funding category they went through.
So **756** Startups were operating which went through just **1** round of funding, and **39** startups had closed down with **1** round of funding



Conclusion:

This project aimed to explore the financial characteristics of Indian startups by examining their operational status in relation to the funds they raised and the number of funding rounds they underwent. Through detailed statistical analysis and visualization, the following key findings emerged:

- **Mean Funds Raised:** The independent sample t-test indicated no statistically significant difference in the mean funds raised between startups that are currently operating and those that have closed. This suggests that the total amount of funding may not be a decisive factor in determining whether a startup succeeds or fails.
- **Funding Rounds:** Similarly, the chi-square test of independence did not show a statistically significant difference in the number of funding rounds between operating and closed startups. This implies that the number of funding rounds alone may not significantly influence a startup's chances of survival.

These findings challenge common assumptions that higher funding amounts and more funding rounds are directly linked to startup success. The results suggest that other factors, possibly beyond just financial metrics, could play a more crucial role in determining the long-term viability of startups.

Future research could benefit from investigating additional factors, such as the startup's business model, market conditions, or the experience of the founding team, to better understand what drives success in the startup ecosystem.