Report

• Data cleaning

Given the time limitations, a decision was made to remove certain column that are less likely relevant for a general analysis of a startup success/failure. The reasons for removing are the following:

1. Columns with as many unique values as the number of rows are not representative

2. It's pointless to have columns with similar context and overlapping meanings

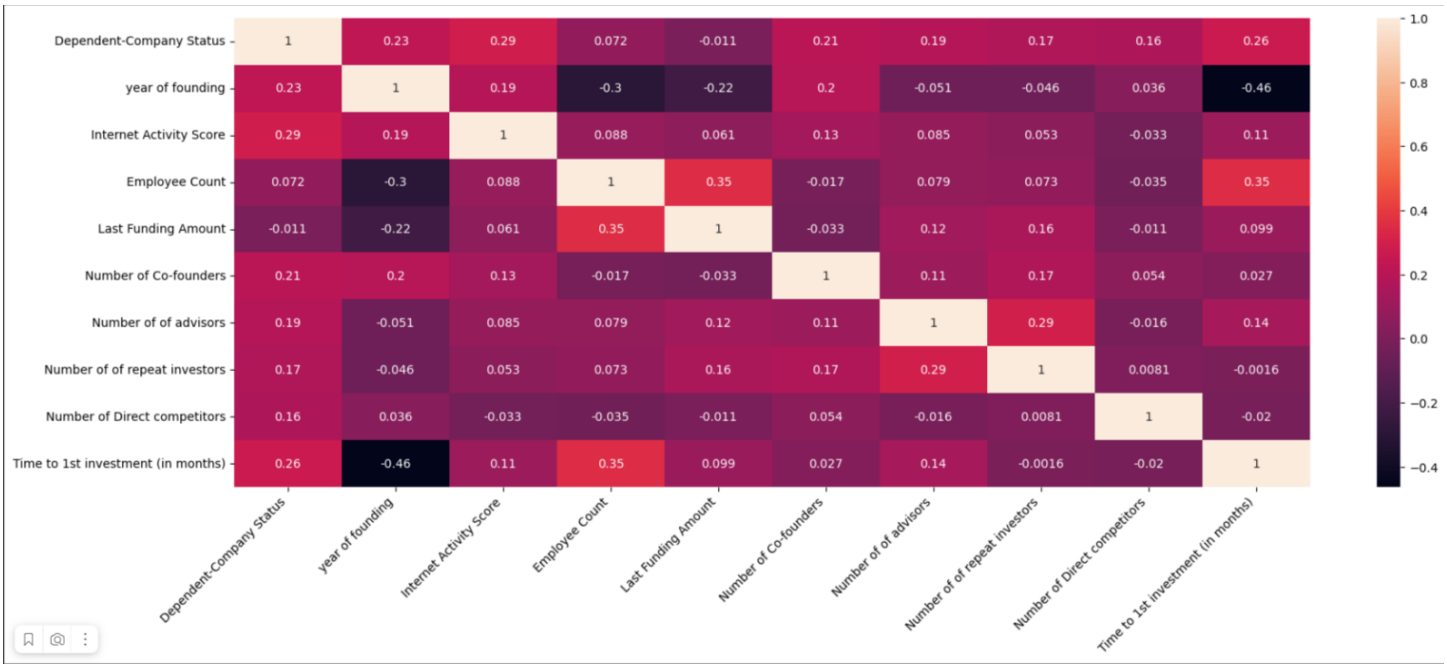
3. With the use of the functions *categorize_and_plot* and *histogram_for_unique_values* it was determined that 'operations' is the most frequent focus function and 'technology' is the most frequent industry. I decided to remove the columns that are not directly related to them.

• Analysis

1. Correlation matrix and heatmap show linear correlation between the numerical columns. It can be seen that the strongest linear correlation is between:

- company status and internet activity score
- company status and time to 1st investment
- a little bit weaker but significant correlation between company status and year of founding

the highest correlation is with internet activity score, which means that the more a company is active in the internet, the more it's visible for customers



2. Point biserial correlation coefficient is useful if we want to find correlations when one of the columns is binary and another is continuous

Correlation with internet activity score: 0.2912979194166299, pvalue: 1.1017834409779038e-10

- a higher Internet Activity Score is associated with a successful company
- the p-value is close to zero, therefore, this relationship is statistically significant

Correlation with investment: 0.2647325758105561, pvalue: 5.196920866888673e-09

- the companies with longer time to 1st investment are likely to be more successful
- the p-value again is very small, so this relationship is statistically significant
- a longer time to 1st investment shows thorough preparation and careful approach to seeking investors

3. Grouping the numerical columns using their means brings us to the following conclusions:

Successful companies have:

- a higher average Internet Activity Score compared to unsuccessful ones
- almost no difference in the last funding amount
- a larger average employee count
- a higher average number of co-founders
- more advisors
- almost no difference in number of repeat investors
- more direct competitors
- a much longer time to first investment
- unsuccessful companies tend to be founded in 2008, this may be due to the crisis of that year

Dependent-Company Status	year of founding	Internet Activity Score	Employee Count	Last Funding Amount	Number of Co-founders	Number of of advisors	Number of of repeat investors	Number of Direct competitors	Time to 1st investment (in months)
0	2008.718563	31.772617	26.329341	5.309543e+06	1.550898	0.502994	0.281437	0.958084	5.550898
1	2009.881967	159.270012	33.967213	5.216624e+06	2.042623	1.298361	0.704918	2.377049	15.281967

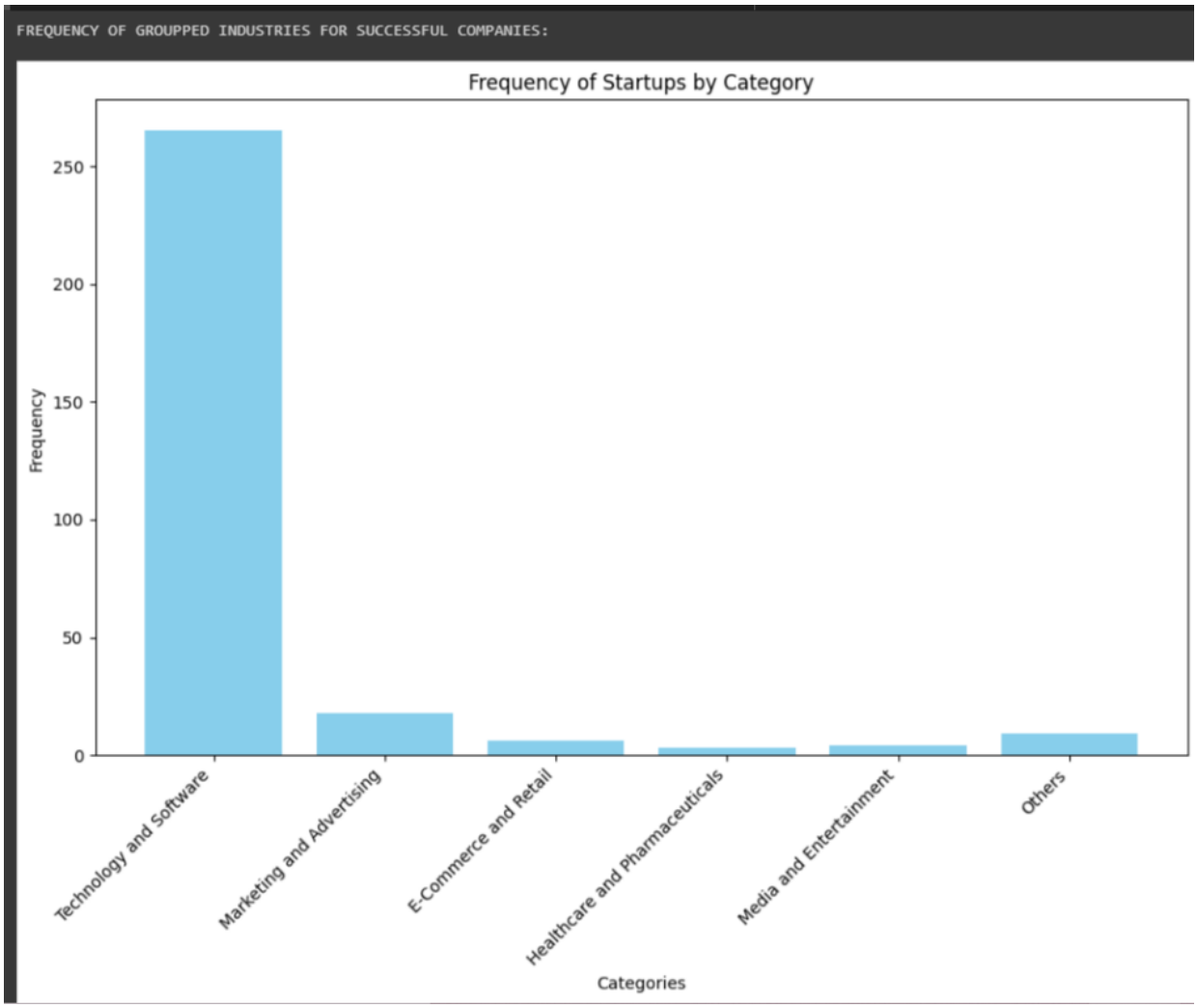
+ Код + Текст

4. Distant correlation finds both linear and non-linear relationships:

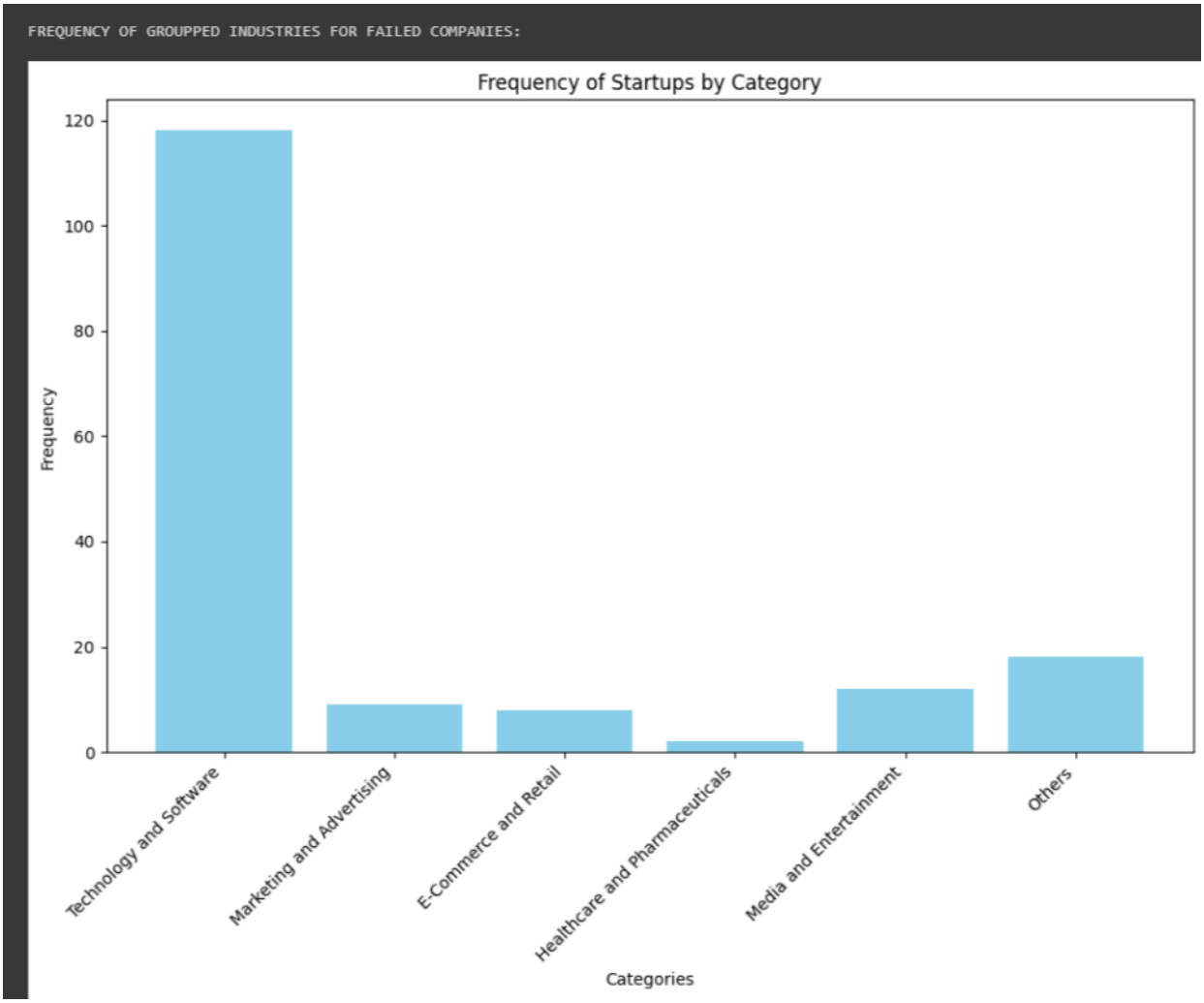
```
Dependent-Company Status 1.0
year of founding 0.2744230632777665
Internet Activity Score 0.34310791705353444
Employee Count 0.1629804052362217
Last Funding Amount 0.09522273694964278
Number of Co-founders 0.2287402805269972
Number of of advisors 0.1929627601410972
Number of of repeat investors 0.19563897129326874
Number of Direct competitors 0.21175708494079118
Time to 1st investment (in months) 0.34284895146111893
```

The most correlated are internet activity score, year of founding, time to 1st investment.

5. Histograms of industries for successful and unsuccessful companies:



FREQUENCY OF GROUPPED INDUSTRIES FOR FAILED COMPANIES:



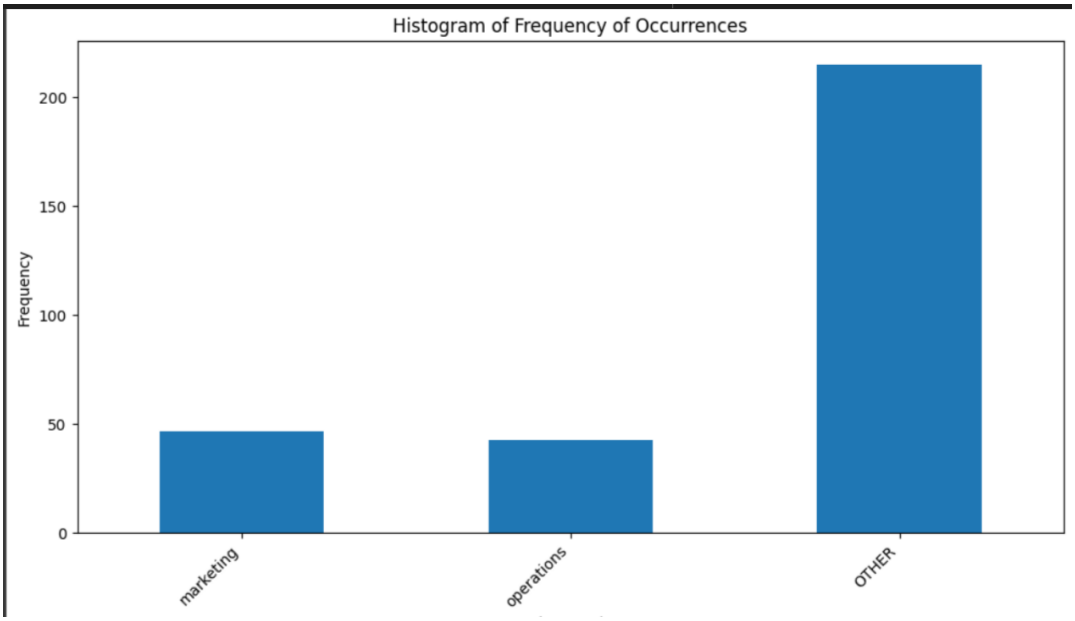
Conclusions:

- the "others" category has a higher frequency among failed startups, which means that companies working in niche industries don't have the same level of support or are likely to face challenges that lead to failure
- in technology and software there're more resources and opportunities available, leading to a higher likelihood of success; at the same time, this category is leading among failed companies too (although with a lower frequency), this can mean that competition in this field is really fierce
- as for e-commerse and retail, it's also highly competitive, and the opportunities for success are limited, due to logistics, market trends, etc.; that's why they're likely to fail rather than be successful
- media and entertainment is leading in failed companies, this may be because monetizing such companies is complicates as users are the source of revenue, and it's difficult to attract them in such a competitive area

6. Histograms of focus functions for successful and unsuccessful companies:

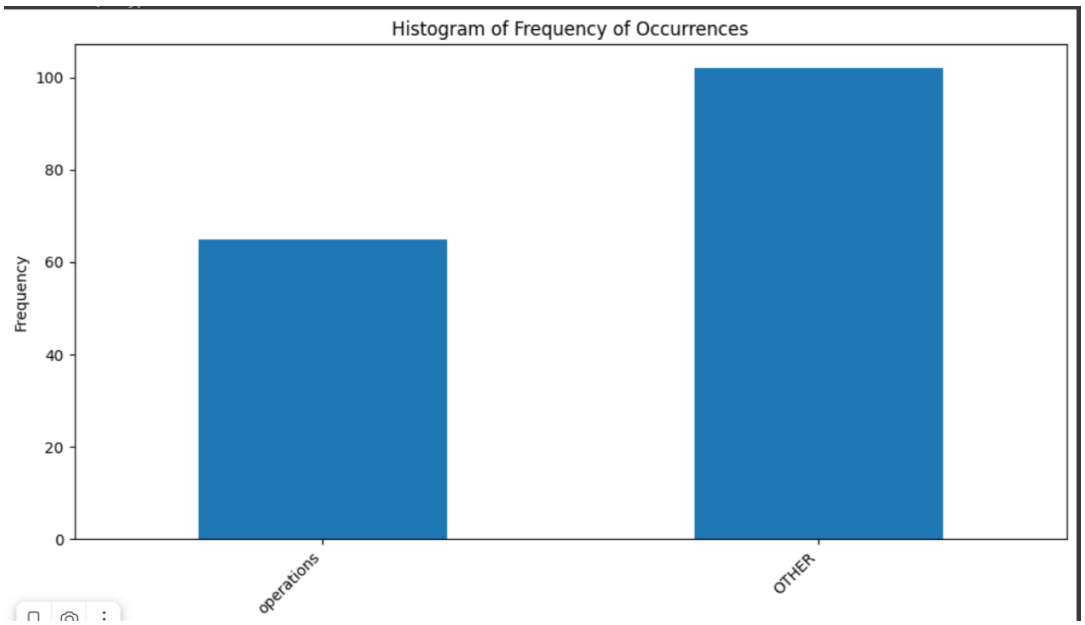
```
FREQUENCY OF GROUPPED FOCUS FUNCTIONS FOR SUCCESSFUL COMPANIES:

Focus functions of company
marketing                                47
operations                               43
analytics                                23
```




```
FREQUENCY OF GROUPPED FOCUS FUNCTION FOR FAILED COMPANIES:

Focus functions of company
operations                               65
marketing                                16
analytics                                10
```



Conclusions:

Marketing is a more frequent focus function among successful companies (47 in the 1st plot vs 16 in the 2nd) because:

- companies with strong marketing often establish better brand recognition and reputation

- companies that understand their market and customers are more likely to be successful
- a well-marketed company can also be more attractive to investors

Operations is a more frequent focus function among failed companies (43 in the 1st plot vs 65 in the 2nd) because:

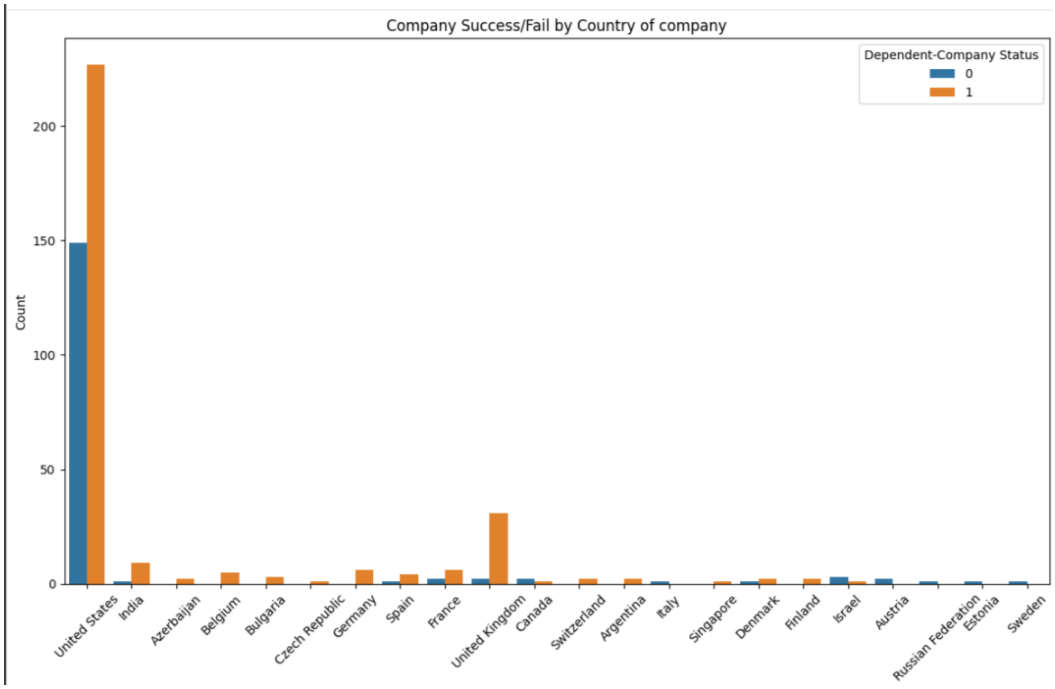
- a heavy focus on operations can indicate that a company is spending too much resources on internal processes
- it can also mean that the companies are facing significant operational challenges at the moment
- a company may have efficient operations but still fail if its products or services do not meet market/customer needs

Analytics is a more frequent focus function among successful companies (23 in the 1st plot vs 10 in the 2nd) because:

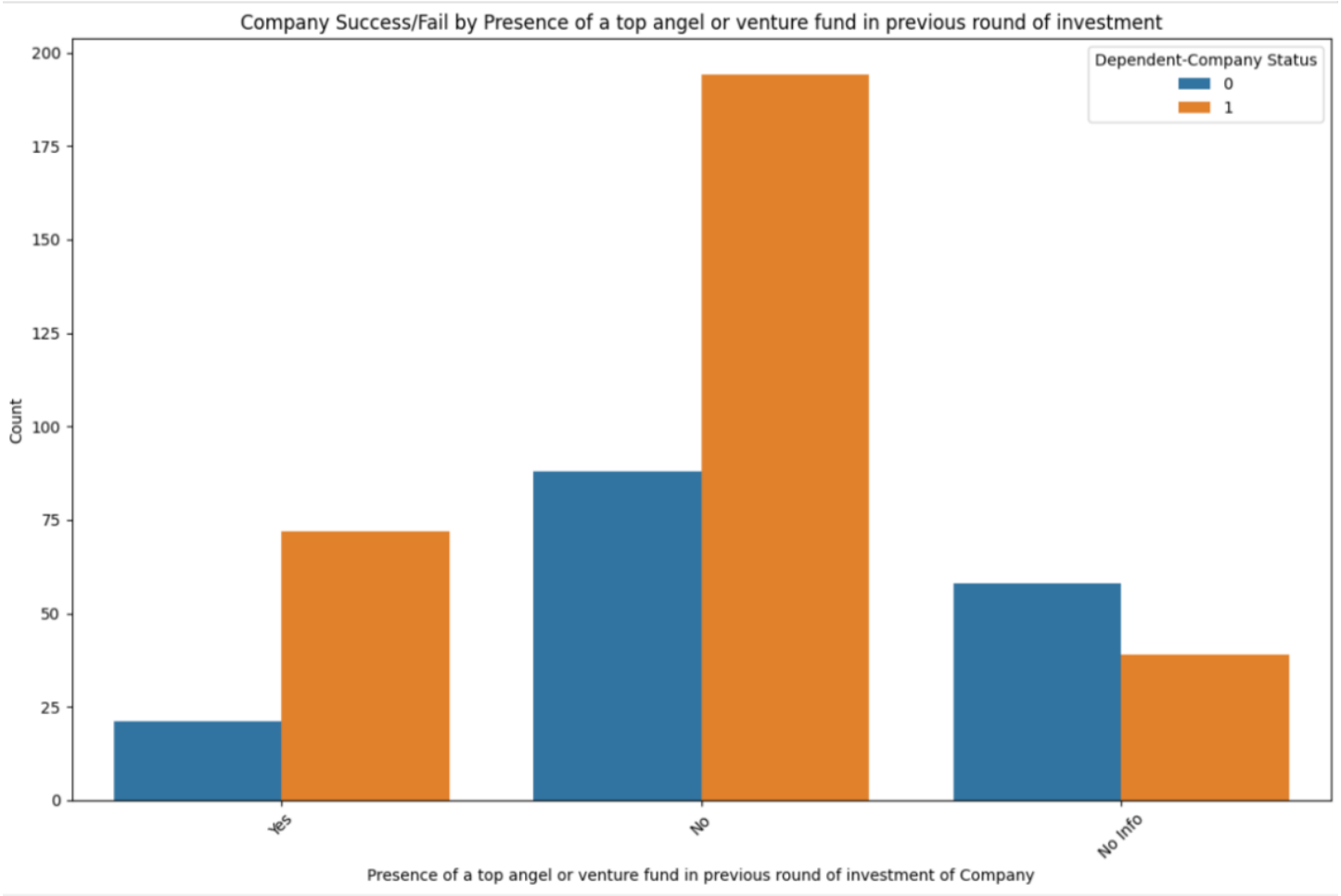
- successful companies use analytics to continuously monitor and optimize their performance
- companies concentrating on analytics really get to know their market, customers, and how well they’re performing
- with analytics they can predict what’s coming, what customers might do, and what could go wrong

7. Countplots for the rest of the categorical data:

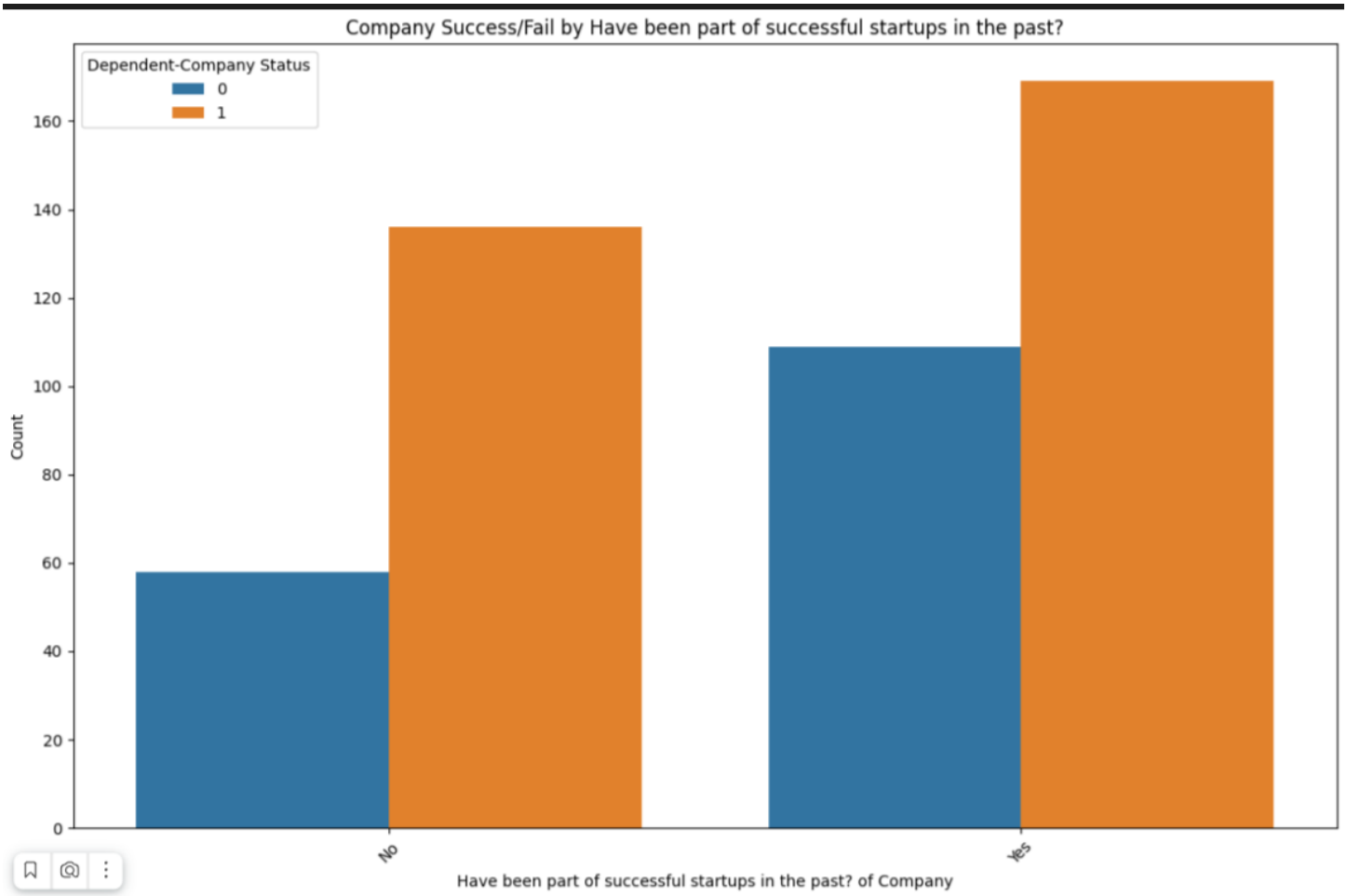
- In the U.S. there're lots of financial resources available for new businesses and a mature entrepreneurship ecosystem. In the UK, there are lots of successful startups, partly because London is a major money center and the government makes it easy for businesses to become successful. The Indian market has a significant potential due to its size and the digital boom, which attracts many entrepreneurs. These are the possible reasons for startups to be successful in these countries.



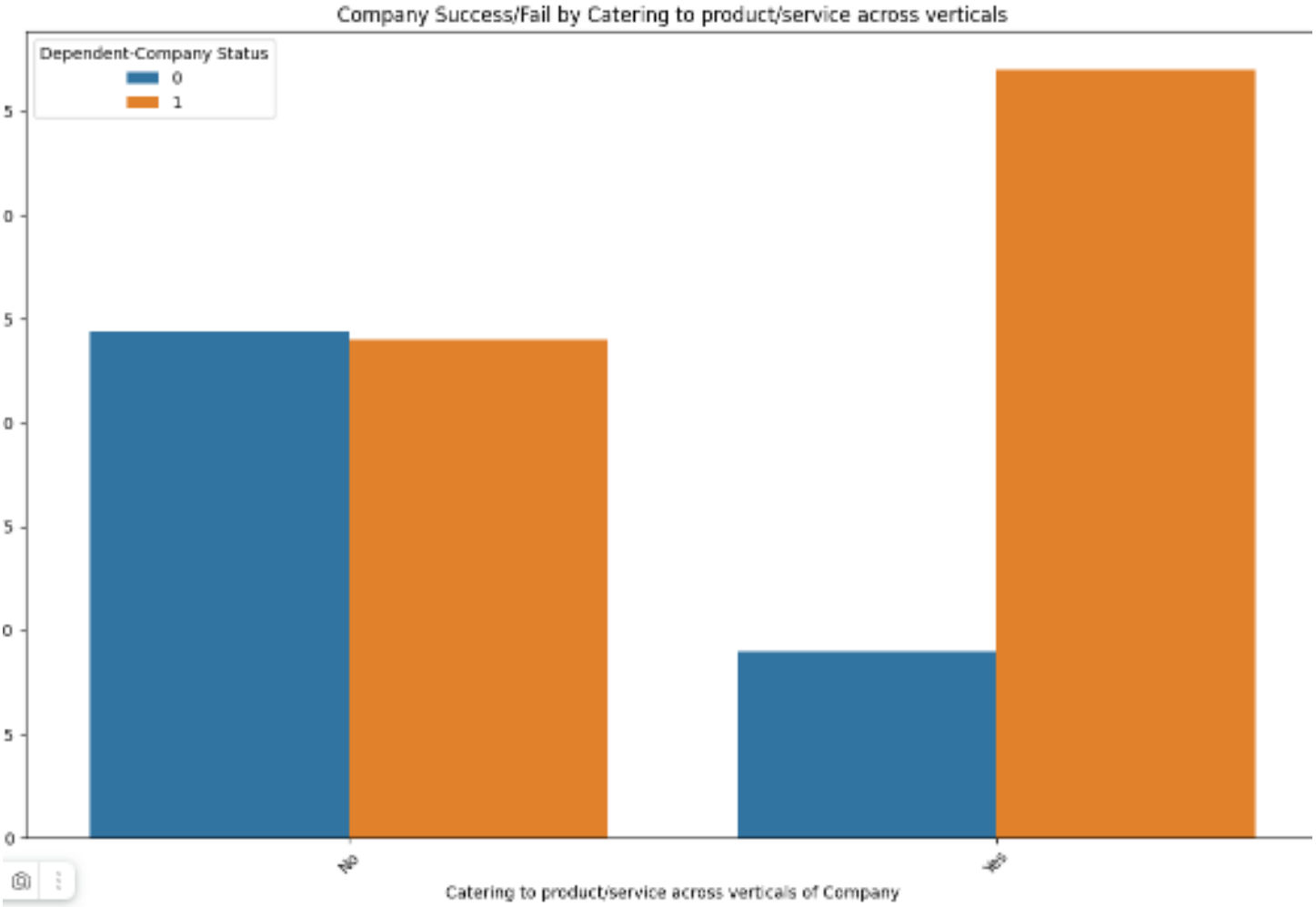
- Companies that had a top angel investor or venture fund involved in the previous funding round have experienced more success than failure. This can suggest that having reputable investors may bring not only money but also valuable networks which can contribute to a company's success. However, there are also a number of successful companies with no information about previous investment from top angels or venture funds, indicating that investment is not the only path to success. It's also possible that companies in the "no info" category could be self-funded or have investments from less well-known investors.



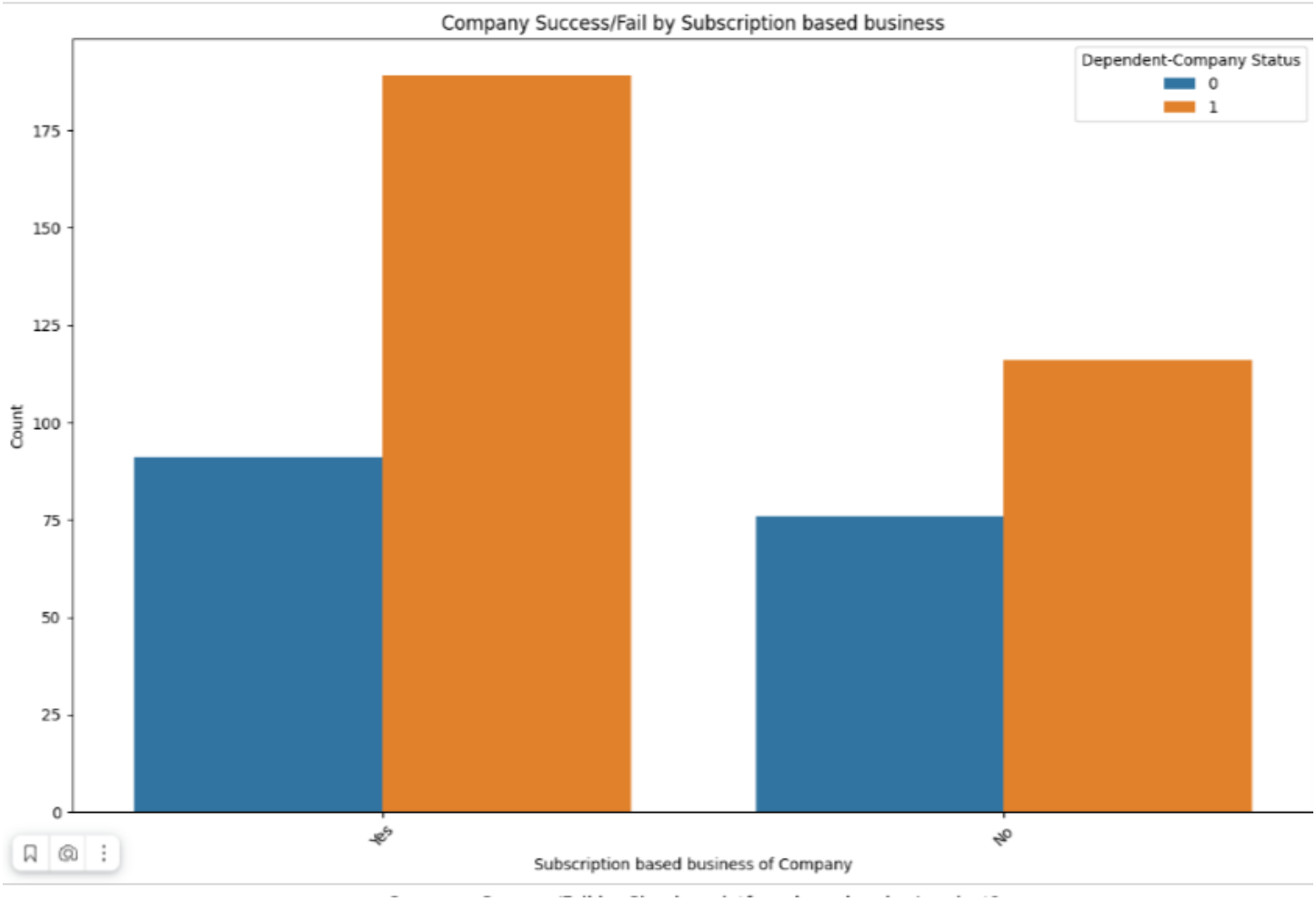
- Companies that have previously been part of successful startups have a higher success rate compared to those that have not. This suggests that experience in successful startups can be an influential factor in the success.



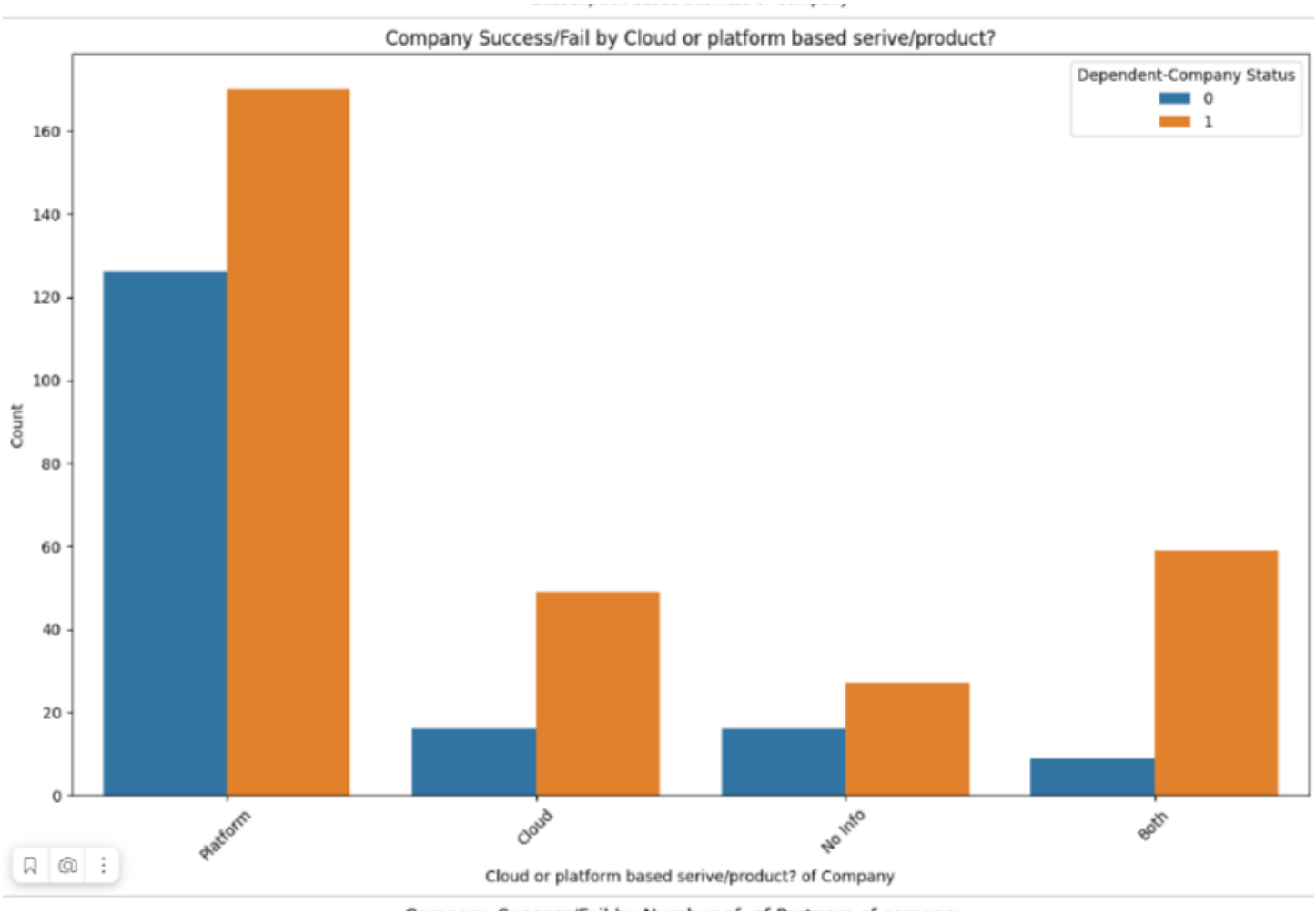
- Product/service oriented companies are tend to be more successful



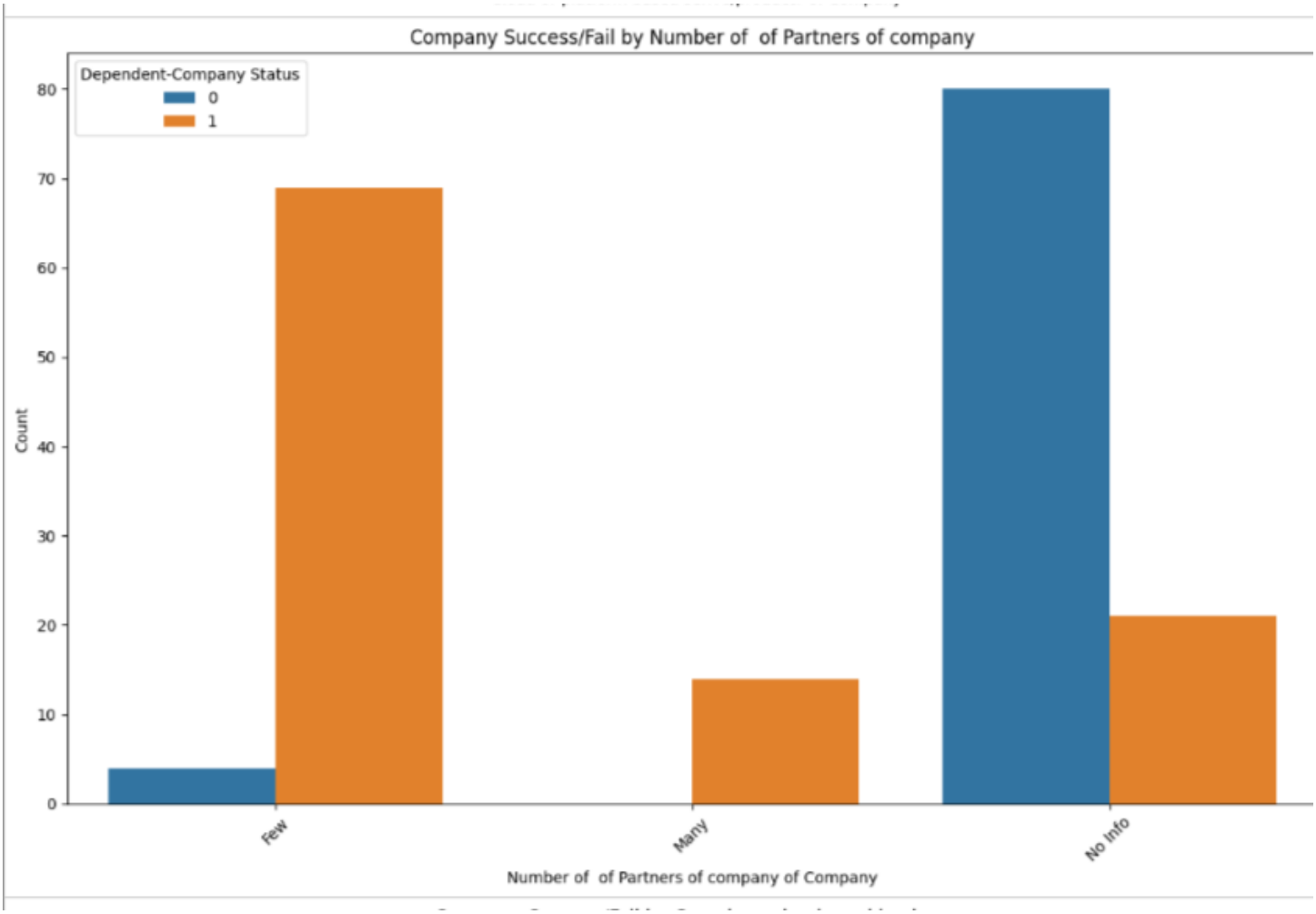
- Subscribtion-based startups are more likely to be successful



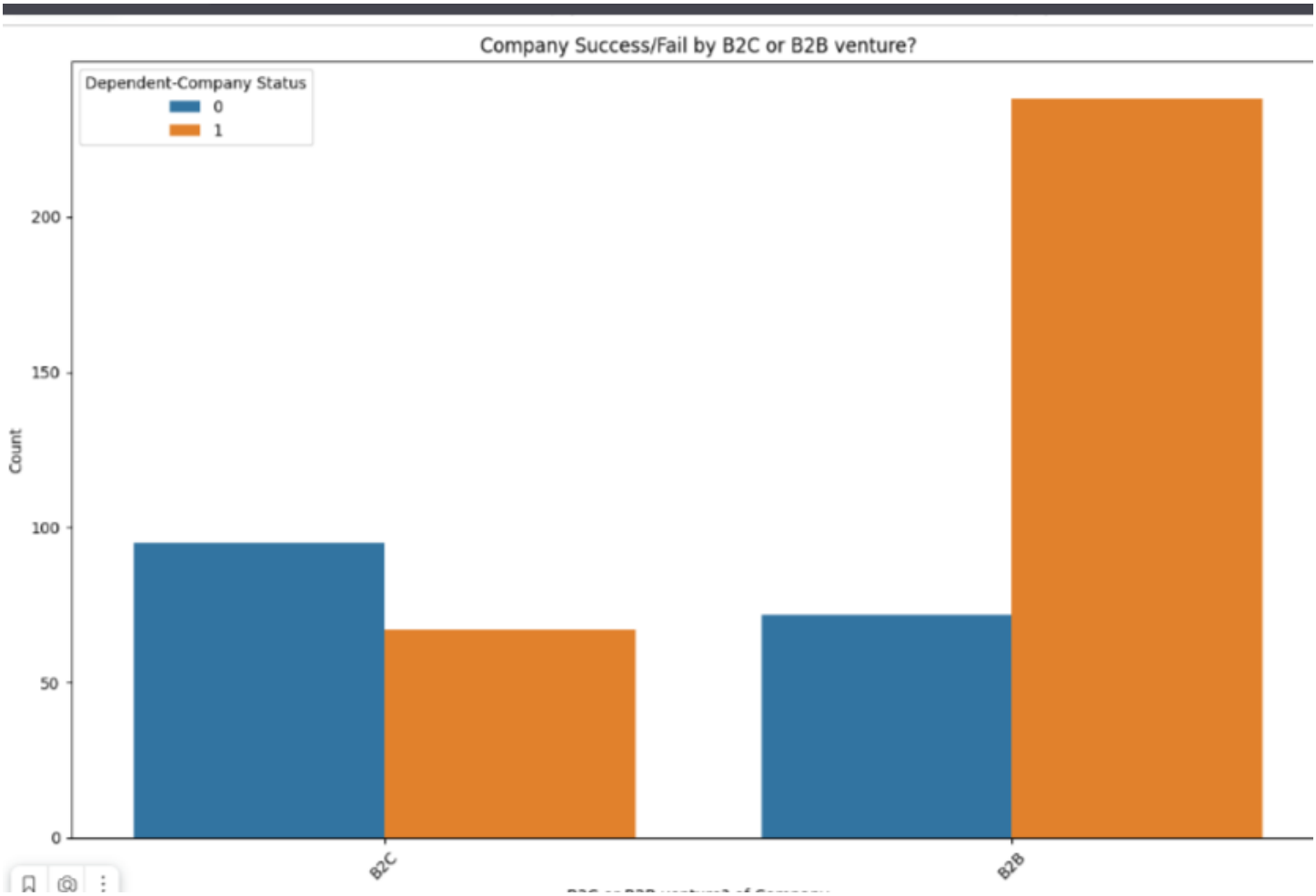
- There seems to be a positive correlation between using a platform and company success. Companies that use both have a greater tendency to being successful. Cloud-based companies seem to struggle more, but this conclusion should be taken with caution as this class is imbalanced. If "No Info" means the companies don’t use either service, then the fact that there’re more successful startups suggests that for some companies not using cloud or platform services doesn’t influence negatively.



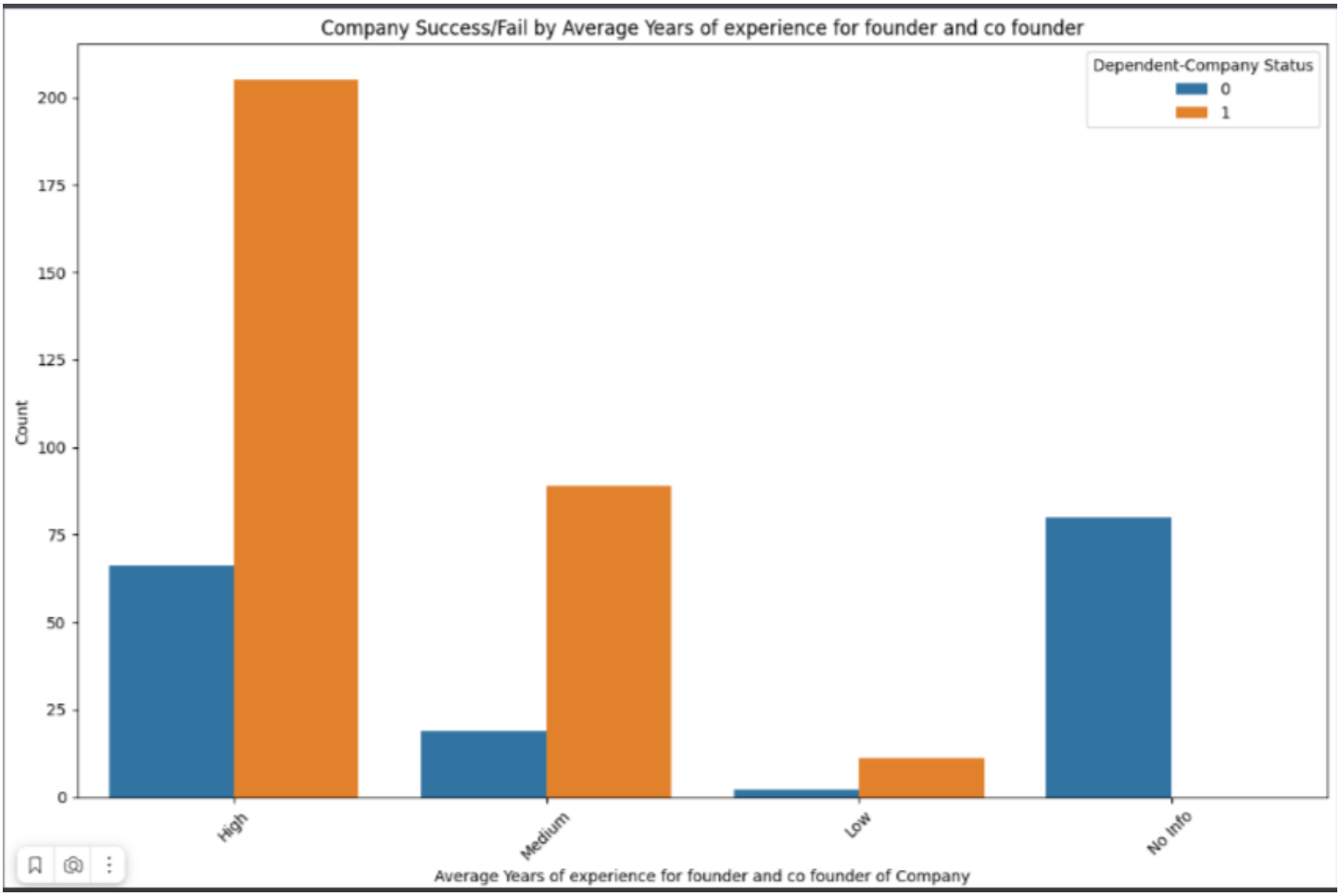
- Companies with few partners have a very high success rate compared to a much smaller failure rate. Companies with many partners are successful. The presence of the "No Info" category can indicate that companies that don't report their number of partners, or where this data isn't available, tend to be less successful because of a lack of transparency which can be a risk factor for failure.



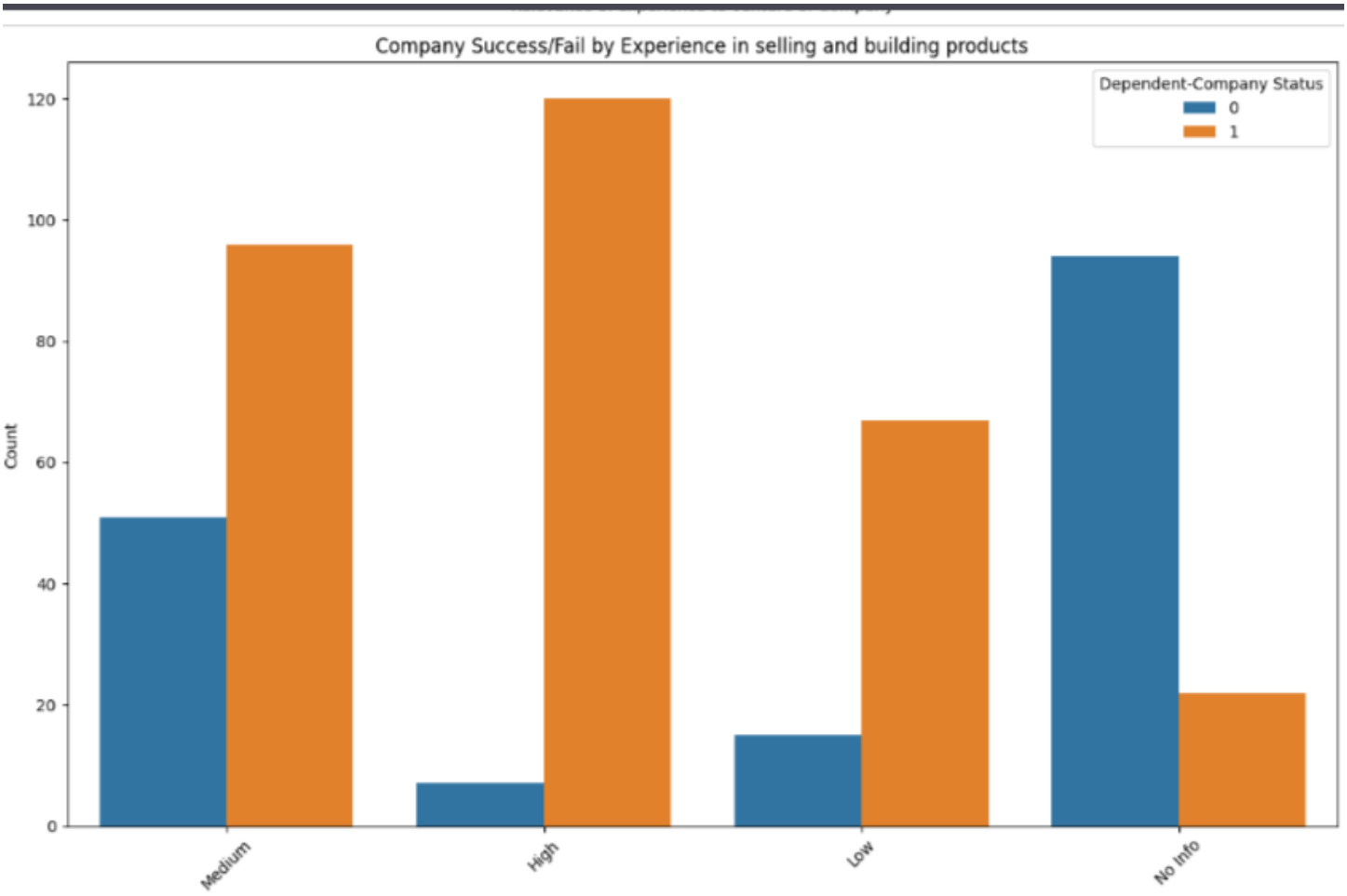
- Businesses selling to other businesses tend to succeed more often than businesses selling to consumers.



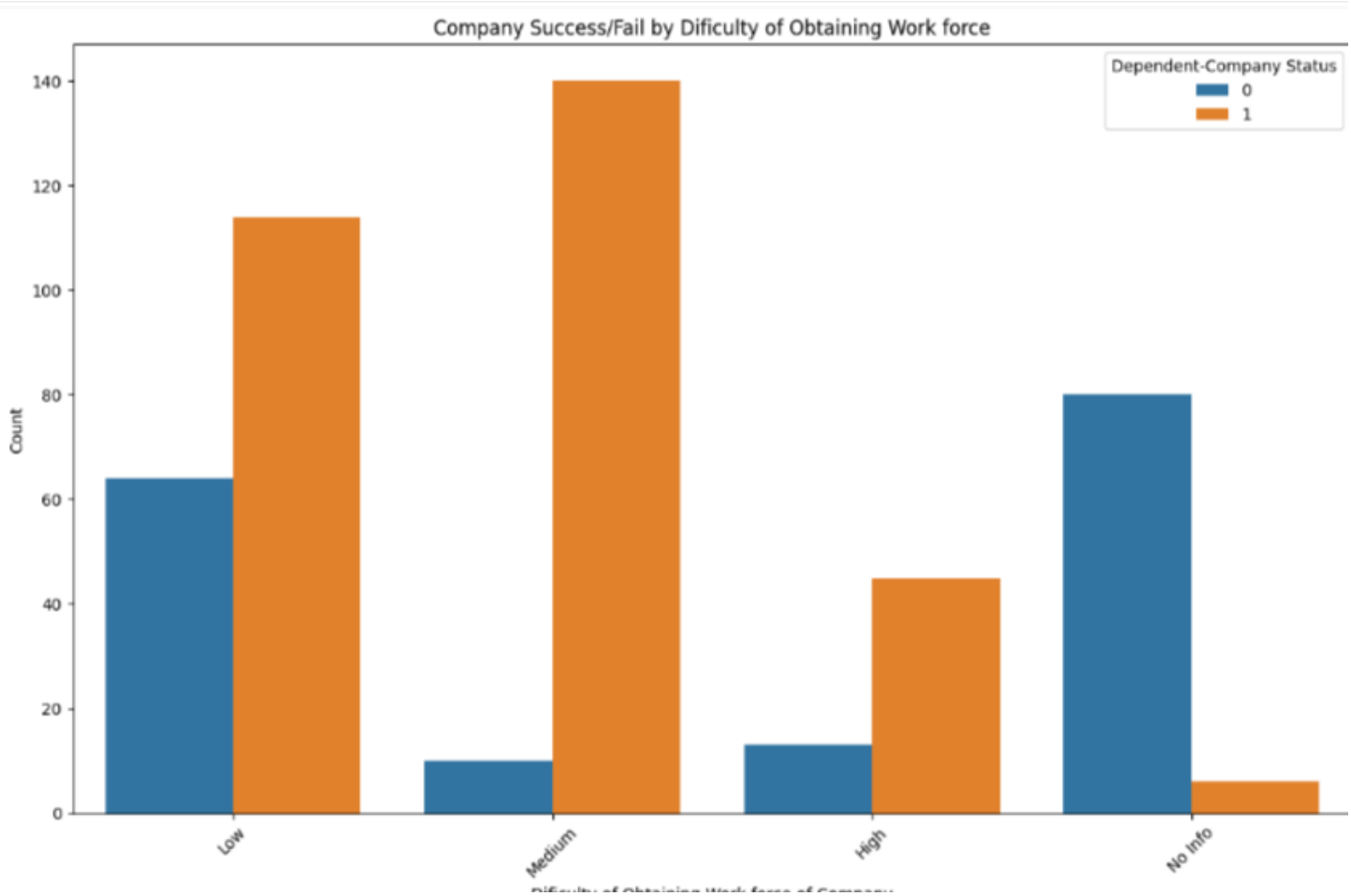
- Companies with more experienced leaders tend to have higher success rates. As for the No Info companies, the absence of information can reflect a lack of transparency, which can be a red flag for investors, customers, and partners, that is why we see high rates of failure.



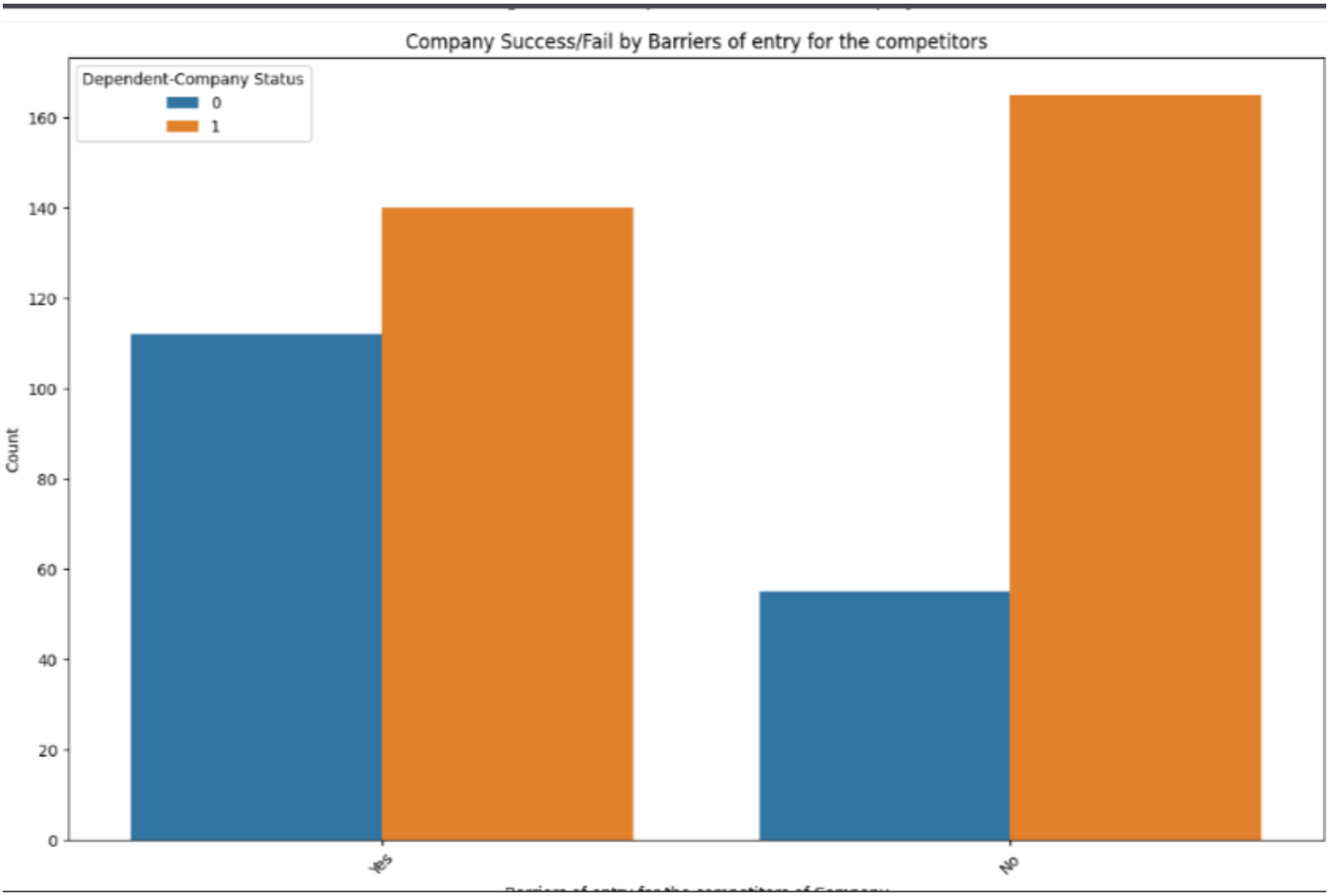
- Expertise in sales is an important factor in company success. Companies with more experience in this area tend to be more successful. The 'No Info' category's high failure rate could suggest that companies that do not conceal this information, or do not have it recorded, can be at a higher risk of failing. Companies that do not provide information may face trust issues.



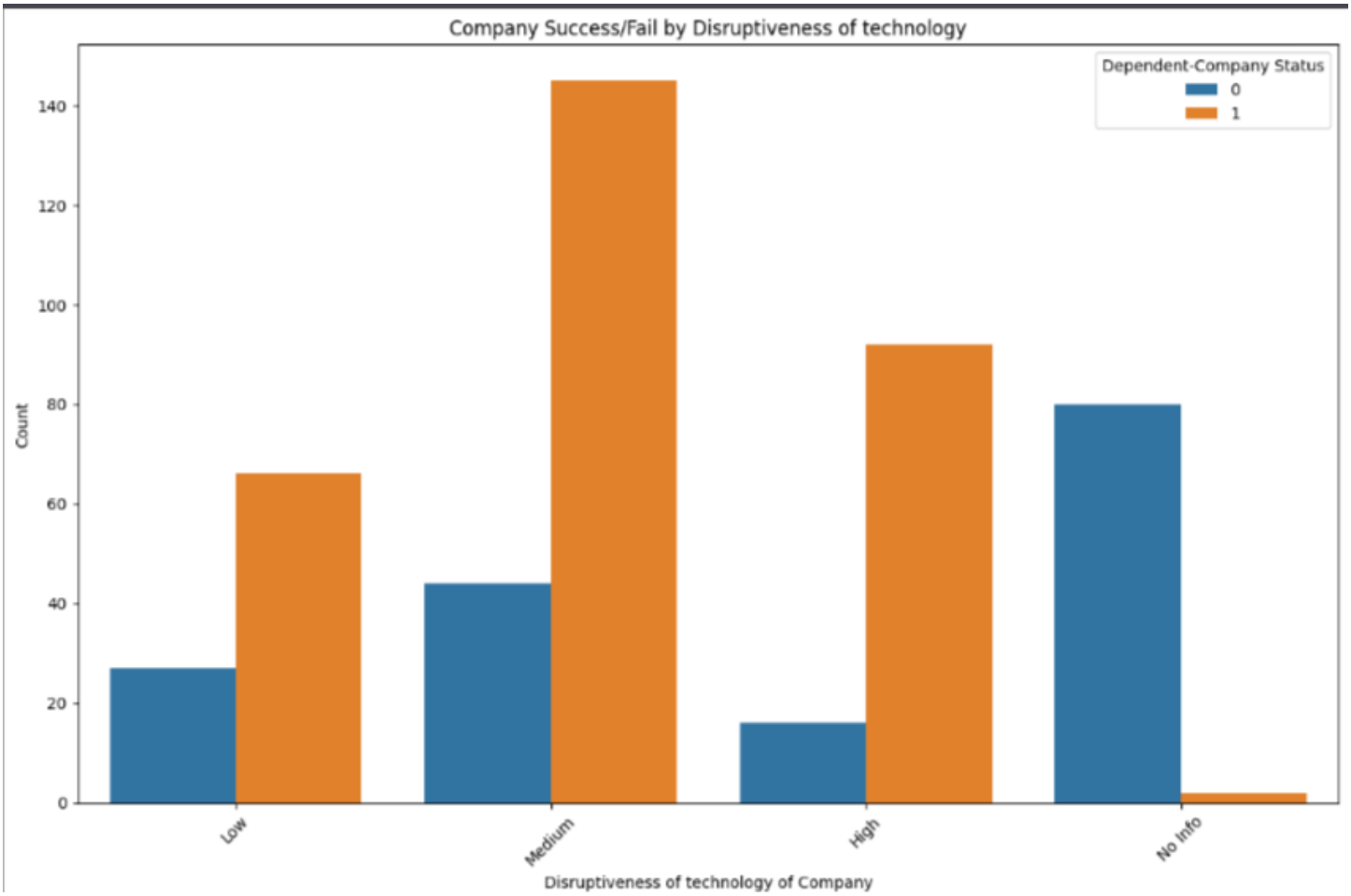
- Companies that find it easier to hire suitable employees tend to be more successful. Reasons for 'No Info' companies being unsuccessful: companies may face internal challenges that make them reluctant to share or acknowledge their hiring difficulties. Also, newer companies can not have enough history to provide this information.



- The stronger are ther barriers, the more difficult it is to enter the market or resist competition



- There seems to be a trend where companies with medium to high disruptiveness of technology are more successful. Regarding No Info companies, companies without information on their technology's disruptiveness might not focus on innovation and it also may reflect a lack of data management or tracking of data. All this leads to failure.


Company Success/Fail by Disruptiveness of technology
Dependent-Company Status
0
1
Count
0
20
40
60
80
100
120
140
Low
Medium
High
No Info
Disruptiveness of technology of Company