Startup Funding in New York City

Yasna Kazimi

2025-06-15

Introduction

New York City is globally recognized as a thriving hub for startups, innovation, and venture capital. With an ecosystem valued at over $621 billion and home to more than 150 unicorns, NYC attracts billions in startup funding and continues to be a leading destination for entrepreneurs and investors alike (Startup Genome). This project explores a critical question within this high-growth environment:

Do Startups in NYC that receive more funding tend to succeed more often?

Using a dataset of startup investment activity in NYC, this analysis focuses on understanding the relationship between funding amount and startup status (e.g., acquired, closed, operating), identifying which sectors receive the most investment, and exploring whether it is possible to predict a startup's success or failure based on its funding and industry.

Importing the Libraries

To support data analysis and visualization throughout this project, several core R packages were loaded. The (tidyverse) package includes essential tools such as dplyr for data manipulation and (ggplot2) for creating visualizations. (janitor) is used to clean column names for easier referencing, while (scales) helps format numerical outputs like funding amounts into readable currency. Lastly, (knitr) is used to produce clean tables and support the final PDF report formatting.

```r
# Core libraries
library(ggplot2)
library(tidyverse)
library(dplyr)
library(janitor)
library(scales)
library(knitr)
```

Importing the Dataset

The dataset used for this project, investment.csv, was imported using the read_csv() function from the readr package, which is part of the tidyverse. This function efficiently loads CSV files into R and automatically parses column types, making it well-suited for data analysis tasks.

```r
# Loading the required dataset
df <- read_csv("investment.csv")
```

Quick Overview of Data

Before beginning the analysis, it’s important to understand the structure and key fields of the dataset. The file investment.csv contains records of startup investments in New York City, including details such as each startup’s funding amount, current status (e.g., acquired, closed, or operating), industry category (category_list), and various funding rounds and dates.

This dataset offers a rich snapshot of NYC’s startup ecosystem and includes companies from a wide range of sectors such as software, health care, e-commerce, and biotechnology. Some startups list multiple sectors within a single record, which may reflect their diverse market strategies. Additionally, the funding amounts vary significantly; from just a few thousand dollars to over a hundred million; which gives us the opportunity to explore whether more funding correlates with a higher chance of success.

Table 1: Preview of NYC Startup Dataset

category_list	funding_total_usd	status
|Entertainment|Politics|Social Media|News|	17,50,000	acquired
|Games|	40,00,000	operating
|Publishing|Education|	40,000	operating
|Electronics|Guides|Coffee|Restaurants|Music|iPhone|Apps|Mobile|iOS|E-Commerce|	15,00,000	operating
|Tourism|Entertainment|Games|	60,000	operating
|Advertising|	49,12,393	closed
|Curated Web|	20,00,000	operating

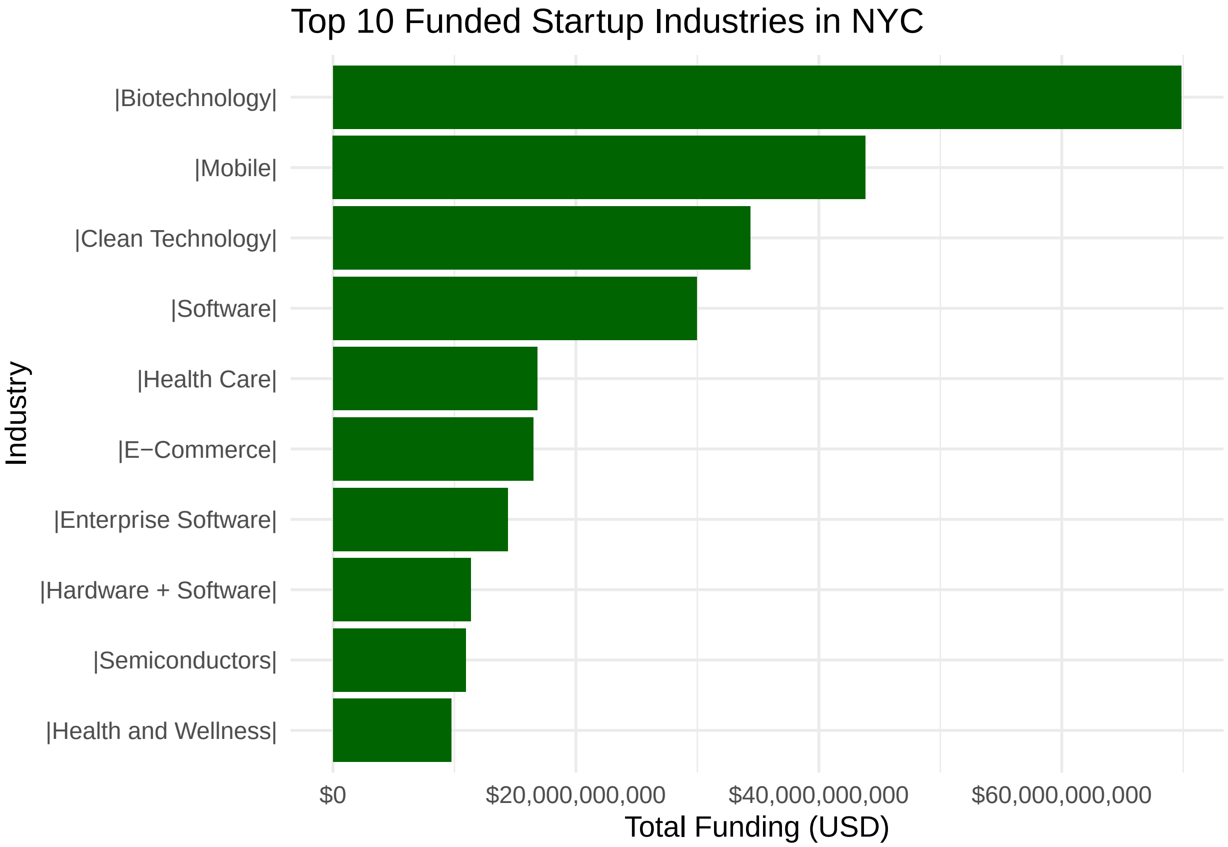
Data Cleaning

To prepare the dataset for analysis, I created a cleaned version called df_clean by removing any rows that had missing or blank values in key columns. Specifically, I filtered out records where the funding amount, startup status, or industry category was either missing (NA) or left empty (“ “). This ensures that the remaining data is complete and reliable for analysis, avoiding errors and improving the accuracy of results in visualizations and machine learning models

```r
df_clean <- df %>% filter(

!is.na(funding_total_usd), funding_total_usd != "", !is.na(status), status != "", !is.na(category_list)
)
```

Exploratory Data Analysis and Data Visualization

```r
df$funding_total_usd <- as.numeric(gsub("[^0-9.]", "", df$funding_total_usd))

top_industries <- df %>%
  filter(!is.na(category_list), category_list != "") %>%
  group_by(category_list) %>%
  summarise(total_funding = sum(funding_total_usd, na.rm = TRUE)) %>%
  arrange(desc(total_funding)) %>%
  slice_max(total_funding, n = 10)

# Plot
ggplot(top_industries, aes(x = reorder(category_list, total_funding), y = total_funding)) +
  geom_col(fill = "darkgreen") +
  coord_flip() +
  scale_y_continuous(labels = scales::dollar) +
  labs(title = "Top 10 Funded Startup Industries in NYC",
       x = "Industry",
       y = "Total Funding (USD)") +
  theme_minimal()
```



According to the chart, Biotechnology stands out as the leading industry with a total funding amount of approximately $70 billion, followed by Mobile with $45 billion and Clean Technology with around $30 billion.

Funding Distribution by Startup Status

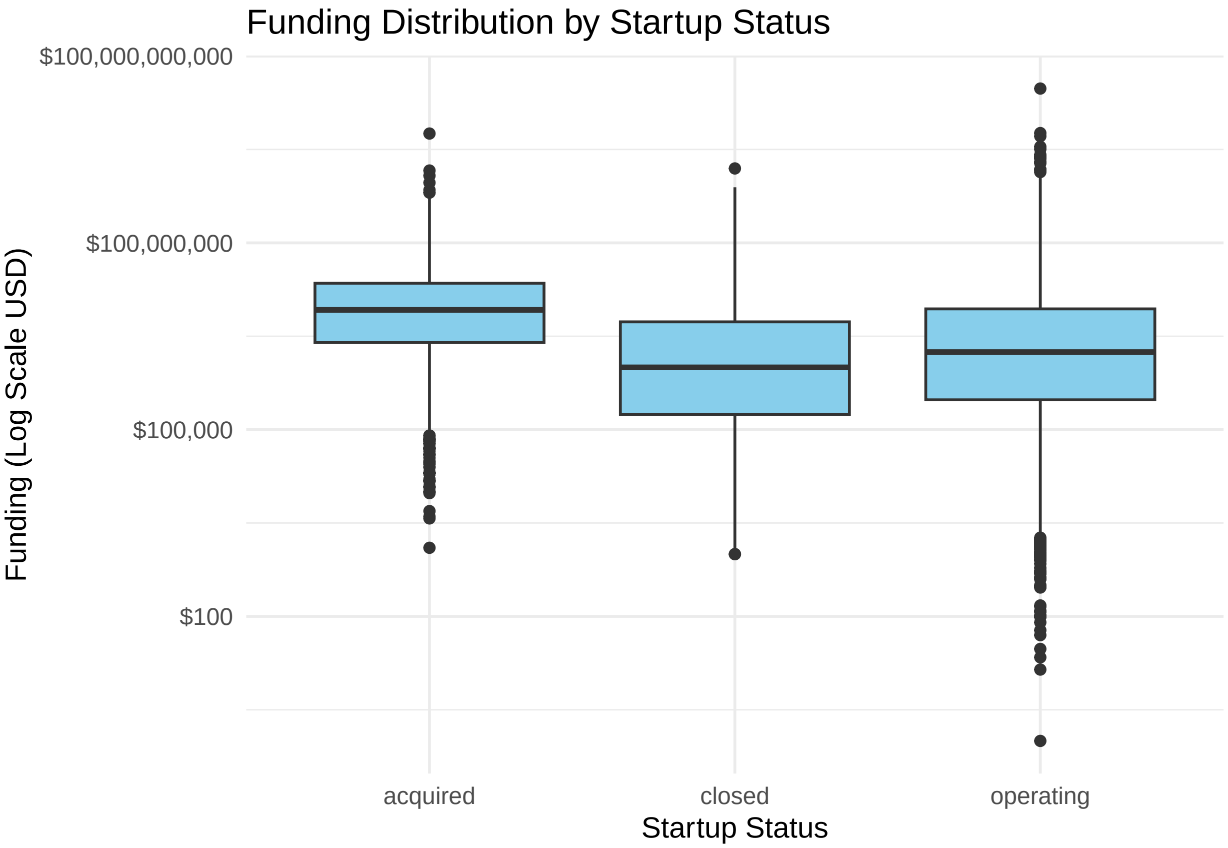
To understand how funding relates to startup outcomes, I looked at the total and average funding received by startups based on their current status—whether they were acquired, still operating, or closed. Startups that were acquired received the most funding on average, about $23.8 million per company. Those still operating had an average of $15.7 million, while startups that closed received the least, around $8.5 million. This pattern suggests that startups with more funding tend to have a better shot at success, especially in achieving an acquisition.

```
df %>%
  filter(!is.na(status), status != "") %>%
  group_by(status) %>%
  summarise(
    count = n(),
    total_funding = sum(funding_total_usd, na.rm = TRUE),
    average_funding = mean(funding_total_usd, na.rm = TRUE)
  ) %>%
  mutate(
    total_funding = scales::dollar(total_funding),
    average_funding = scales::dollar(average_funding)
  )
```

```
## # A tibble: 3 x 4
##   status    count total_funding    average_funding
##   <chr>     <int> <chr>            <chr>
## 1 acquired   3692 $76,630,346,870  $23,813,035
## 2 closed     2603 $18,281,373,291  $8,471,443
## 3 operating 41829 $541,787,796,768 $15,737,750
```

The Graph

```
df %>%
  filter(!is.na(status), status != "") %>%
  ggplot(aes(x = status, y = funding_total_usd)) +
  geom_boxplot(fill = "skyblue") +
  scale_y_log10(labels = scales::dollar) +
  labs(
    title = "Funding Distribution by Startup Status",
    x = "Startup Status",
    y = "Funding (Log Scale USD)"
  ) +
  theme_minimal()
```

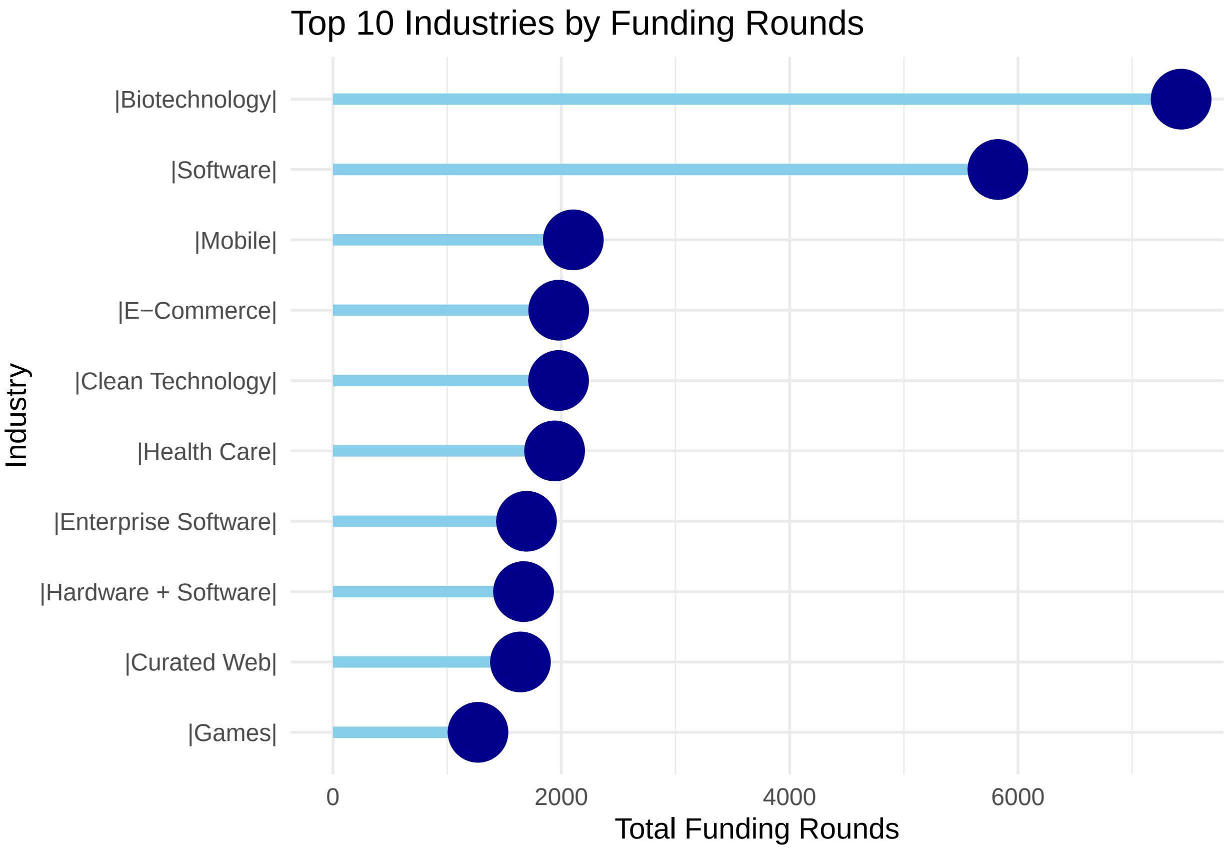


The boxplot compares funding distributions across startup statuses. Acquired startups generally received higher funding amounts, while closed startups had lower medians. This supports the idea that greater funding may be linked to higher chances of acquisition.

Top Industries with most total Funding Rounds

This lollipop chart highlights the top ten industries in New York City with the highest number of total funding rounds. Biotechnology leads significantly, followed by Software, indicating strong investor interest in these sectors. Mobile, E-Commerce, and Clean Technology also show notable funding activity, reflecting the diversity of innovation and capital attraction across different segments of the startup ecosystem. This distribution suggests that certain industries consistently attract repeat investment, possibly due to high growth potential or proven market demand.

```r
library(ggplot2)
library(dplyr)

top_industries <- df %>%
  filter(!is.na(category_list), category_list != "", !is.na(funding_rounds)) %>%
  group_by(category_list) %>%
  summarise(total_rounds = sum(funding_rounds)) %>%
  arrange(desc(total_rounds)) %>%
  slice_max(total_rounds, n = 10)

ggplot(top_industries, aes(x = reorder(category_list, total_rounds), y = total_rounds)) +
  geom_segment(aes(xend = category_list, y = 0, yend = total_rounds), color = "skyblue", size = 2) +
  geom_point(color = "darkblue", size = 10) +
  coord_flip() +
  labs(
    title = "Top 10 Industries by Funding Rounds",
    x = "Industry",
    y = "Total Funding Rounds"
  ) +
  theme_minimal()
```



Now let's answer the main question: Does more funding lead to more success for startup companies? The answer is yes—startups that receive higher amounts of funding are generally more likely to succeed. The more financial support they have, the greater their chances of achieving positive outcomes like acquisition.

Can Funding Amount and industry be used to predict whether a startup will succeed or fail?

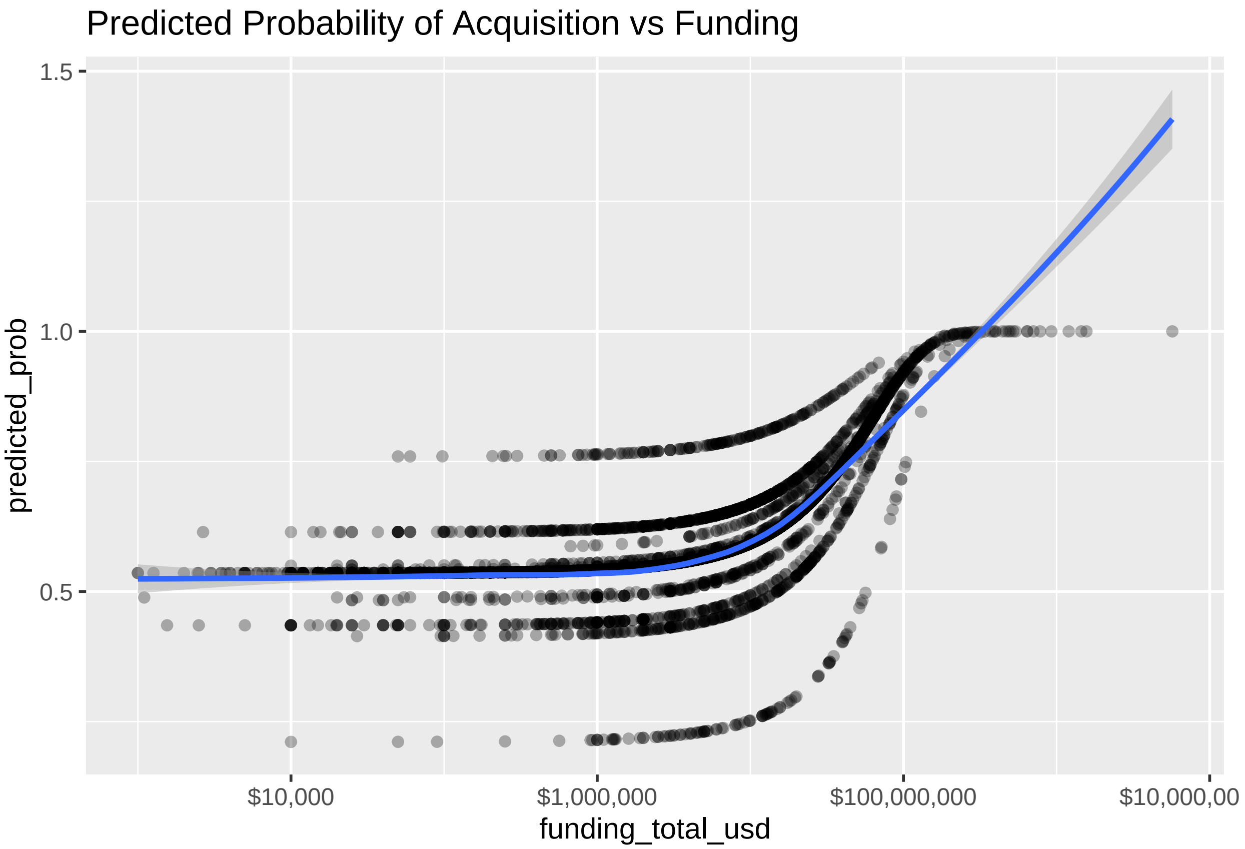
I will use machine learning to evaluate whether Startup success is predictable based on funding and industry — and if so, identify which sectors are most strongly associated with successful outcomes.

```
ml_df <- df %>%
  filter(status %in% c("acquired", "closed"),
         !is.na(funding_total_usd),
         !is.na(category_list)) %>%
  mutate(
    # Convert status to binary: acquired = 1, closed = 0
    success = ifelse(status == "acquired", 1, 0)
  )
#Will take the top 10 sectors and label all others as "Other":
top_sectors <- ml_df %>%
  count(category_list, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  pull(category_list)

ml_df <- ml_df %>%
  mutate(category_grouped = ifelse(category_list %in% top_sectors, category_list, "Other"))

# Category to factor
ml_df$category_grouped <- as.factor(ml_df$category_grouped)

# Fit logistic regression
model <- glm(success ~ funding_total_usd + category_grouped, data = ml_df, family = "binomial")

summary(model)
```

```
##
## Call:
## glm(formula = success ~ funding_total_usd + category_grouped,
##     family = "binomial", data = ml_df)
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            1.688e-01  1.878e-01   0.899  0.36865
## funding_total_usd                      2.317e-08  1.862e-09  12.446  < 2e-16
## category_grouped|Biotechnology|       -5.159e-01  2.249e-01  -2.294  0.02178
## category_grouped|Clean Technology|    -1.488e+00  2.876e-01  -5.175 2.28e-07
## category_grouped|Curated Web|         -4.303e-01  2.296e-01  -1.874  0.06090
## category_grouped|Enterprise Software|  9.800e-01  2.704e-01   3.624  0.00029
## category_grouped|Games|               -2.362e-01  2.655e-01  -0.890  0.37353
## category_grouped|Hardware + Software| -2.147e-01  2.658e-01  -0.808  0.41916
## category_grouped|Mobile|               3.038e-02  2.466e-01   0.123  0.90192
## category_grouped|Semiconductors|       1.673e-01  2.808e-01   0.596  0.55131
## category_grouped|Software|             2.960e-01  2.086e-01   1.419  0.15579
## category_groupedOther                 -2.625e-02  1.906e-01  -0.138  0.89046
##
## (Intercept)
## funding_total_usd                     ***
## category_grouped|Biotechnology|       *
## category_grouped|Clean Technology|    ***
## category_grouped|Curated Web|          .
## category_grouped|Enterprise Software| ***
## category_grouped|Games|
## category_grouped|Hardware + Software|
## category_grouped|Mobile|
## category_grouped|Semiconductors|
## category_grouped|Software|
## category_groupedOther
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7080.3  on 5260  degrees of freedom
## Residual deviance: 6721.8  on 5249  degrees of freedom
## AIC: 6745.8
##
## Number of Fisher Scoring iterations: 7
```

Prediction Graph

```
ml_df$predicted_prob <- predict(model, newdata = ml_df, type = "response")


ggplot(ml_df, aes(x = funding_total_usd, y = predicted_prob)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess") +
  scale_x_log10(labels = dollar_format()) +
  labs(title = "Predicted Probability of Acquisition vs Funding")
```


Predicted Probability of Acquisition vs Funding

While the probability remains relatively stable for startups with lower funding, the LOESS curve reveals a clear upward trend for those receiving higher funding amounts.

Model Summary

To explore whether startup success (defined as acquisition vs. failure) can be predicted using funding and sector, a logistic regression model was applied. The results show that higher funding significantly increases the likelihood of acquisition, with a strong positive correlation between funding amount and success. Additionally, the sector of the startup plays an important role: **startups in the Enterprise Software category were significantly more likely to be acquired, while those in Clean Technology and Biotechnology were less likely to achieve acquisition, even with comparable funding levels.** These findings suggest that while funding is a strong predictor of success, industry-specific factors also influence a startup's likelihood of reaching a successful exit.

Conclusion

This project explored the relationship between startup funding, industry sector, funding rounds, and success outcomes within the New York City startup ecosystem. Through exploratory data analysis and visualizations, we identified that industries such as Biotechnology, Mobile, and Clean Technology received the highest total funding.

- Startups that were acquired had the most funding on average, which shows that more money often leads to better chances of success.
- Companies in sectors like Enterprise Software were more likely to get acquired, while those in Clean Technology and Biotechnology were less likely—even if they had a lot of funding.
- When we looked at funding rounds, we saw that startups with more rounds of investment were usually more likely to be acquired. This suggests that steady support from investors can be a good sign of future success.
- On the other hand, startups with fewer funding rounds were more likely to shut down or stop growing.

Overall, the analysis demonstrates that while funding amount is a key factor in predicting startup success, both the industry and the frequency of funding events play a significant role in shaping startup outcomes in competitive markets like NYC.

References:

https://startupgenome.com/ecosystems/new-york-city