Chasing Unicorns: Valuations, Investors, Insights

A unicorn company is a privately held company with a current valuation of over $1 billion USD. This dataset consists of unicorn companies and startups across the globe as of November 2021, including country of origin, sector, select investors, and valuation of each unicorn.

Note former unicorn companies that have since exited due to IPO or acquisitions are not included in this list.

Scenario

You have been hired as a data scientist for a company that invests in start-ups. Your manager is interested in whether it is possible to predict whether a company reaches a valuation over 5 billion based on characteristics such as its country of origin, its category, and details about its investors.

Using the dataset provided, you have been asked to test whether such predictions are possible, and the confidence one can have in the results.

Data Preparation

```r
# import libraries
suppressPackageStartupMessages(library(tidyverse))
library(ggplot2)
```

```
# import datasets
companies <- read_csv("companies.csv")
```

```
## Rows: 1074 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): company, city, country, continent
## dbl (1): company_id
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
industries <- read_csv("industries.csv")
```

```
## Rows: 1074 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): industry
## dbl (1): company_id
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
funding <- read_csv("funding.csv")
```

```
## Rows: 1074 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): select_investors
## dbl (3): company_id, valuation, funding
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# view datasets
glimpse(companies)
```

```
## Rows: 1,074
## Columns: 5
## $ company_id <dbl> 189, 848, 556, 999, 396, 931, 364, 732, 906, 72, 983, 898, ~
## $ company    <chr> "Otto Bock HealthCare", "Matrixport", "Cloudinary", "PLACE"~
## $ city       <chr> "Duderstadt", NA, "Santa Clara", "Bellingham", "New York", ~
## $ country    <chr> "Germany", "Singapore", "United States", "United States", "~
## $ continent  <chr> "Europe", "Asia", "North America", "North America", "North ~
```

```
glimpse(industries)
```

```
## Rows: 1,074
## Columns: 2
## $ company_id <dbl> 189, 848, 556, 999, 396, 931, 364, 732, 906, 72, 983, 898, ~
## $ industry   <chr> "Health", "Fintech", "Internet software & services", "Inter~
```

```
glimpse(funding)
```

```
## Rows: 1,074
## Columns: 4
## $ company_id       <dbl> 189, 848, 556, 999, 396, 931, 364, 732, 906, 72, 983,~
## $ valuation        <dbl> 4e+09, 1e+09, 2e+09, 1e+09, 2e+09, 1e+09, 2e+09, 1e+0~
## $ funding          <dbl> 0.00e+00, 1.00e+08, 1.00e+08, 1.00e+08, 1.00e+08, 1.0~
## $ select_investors <chr> "EQT Partners", "\"Dragonfly Captial, Qiming Venture ~
```

```
# join the three datasets together
unicorn <- left_join(companies, industries, by= "company_id") %>%
  left_join(funding, by="company_id")
```

The dataset consists of 1024 rows with 9 columns

```
# check for missing values
unicorn %>%
  is.na() %>%
  colSums()
```

```
##       company_id          company             city          country
##                0                0               16                0
##        continent         industry        valuation          funding
##                0                0                0                0
## select_investors
##                1
```

```
# check for duplicate entries
unicorn %>%
  distinct()
```

```
## # A tibble: 1,074 x 9
##    company_id company         city  country continent industry valuation funding
##         <dbl> <chr>           <chr> <chr>   <chr>     <chr>        <dbl>   <dbl>
##  1        189 Otto Bock Heal~ Dude~ Germany Europe    Health         4e9  0
##  2        848 Matrixport      <NA>  Singap~ Asia      Fintech        1e9  1   e8
##  3        556 Cloudinary      Sant~ United~ North Am~ Interne~       2e9  1   e8
##  4        999 PLACE           Bell~ United~ North Am~ Interne~       1e9  1   e8
##  5        396 candy.com       New ~ United~ North Am~ Fintech        2e9  1   e8
##  6        931 HAYDON          Shan~ China   Asia      Consume~       1e9  1   e8
##  7        364 eDaili          Shan~ China   Asia      E-comme~       2e9  1.01e8
##  8        732 CoinTracker     San ~ United~ North Am~ Fintech        1e9  1.02e8
##  9        906 EcoFlow         Shen~ China   Asia      Hardware       1e9  1.05e8
## 10         72 DJI Innovations Shen~ China   Asia      Hardware       8e9  1.05e8
## # i 1,064 more rows
## # i 1 more variable: select_investors <chr>
```

Most missing values are contained in the city column which is not useful for my analysis, so I'll drop the city and company_id columns. The dataset also contains no duplicated entries,

```
unicorn <- unicorn %>%
  select(-city, -company_id)

# describe the dataset
n_distinct(unicorn$company)
```

```
## [1] 1073
```

```
n_distinct(unicorn$country)
```

```
## [1] 46
```

```
n_distinct(unicorn$industry)
```

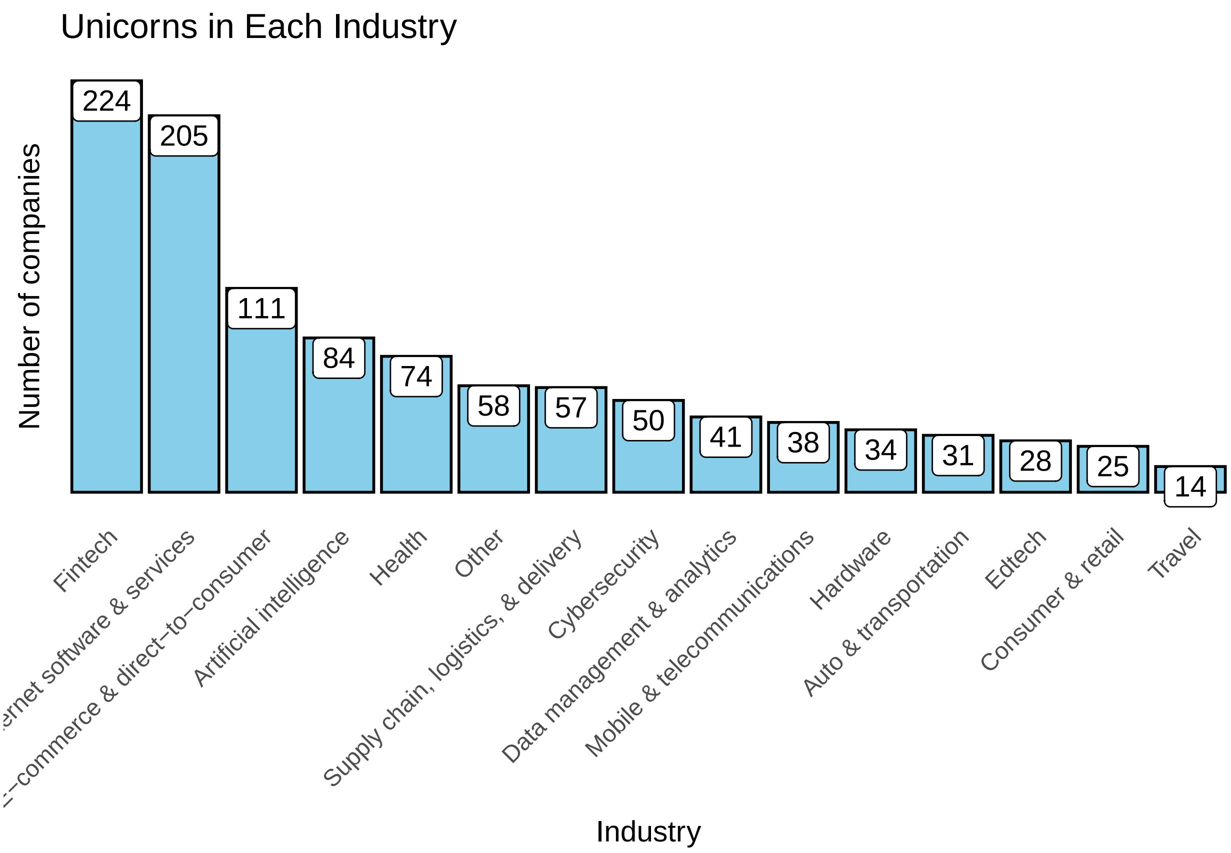
```
## [1] 15
```

This dataset contains records of 1073 unicorn companies from 46 countries of the world, spread out over 15 different industries.

The data has been prepared and is now ready for analysis.

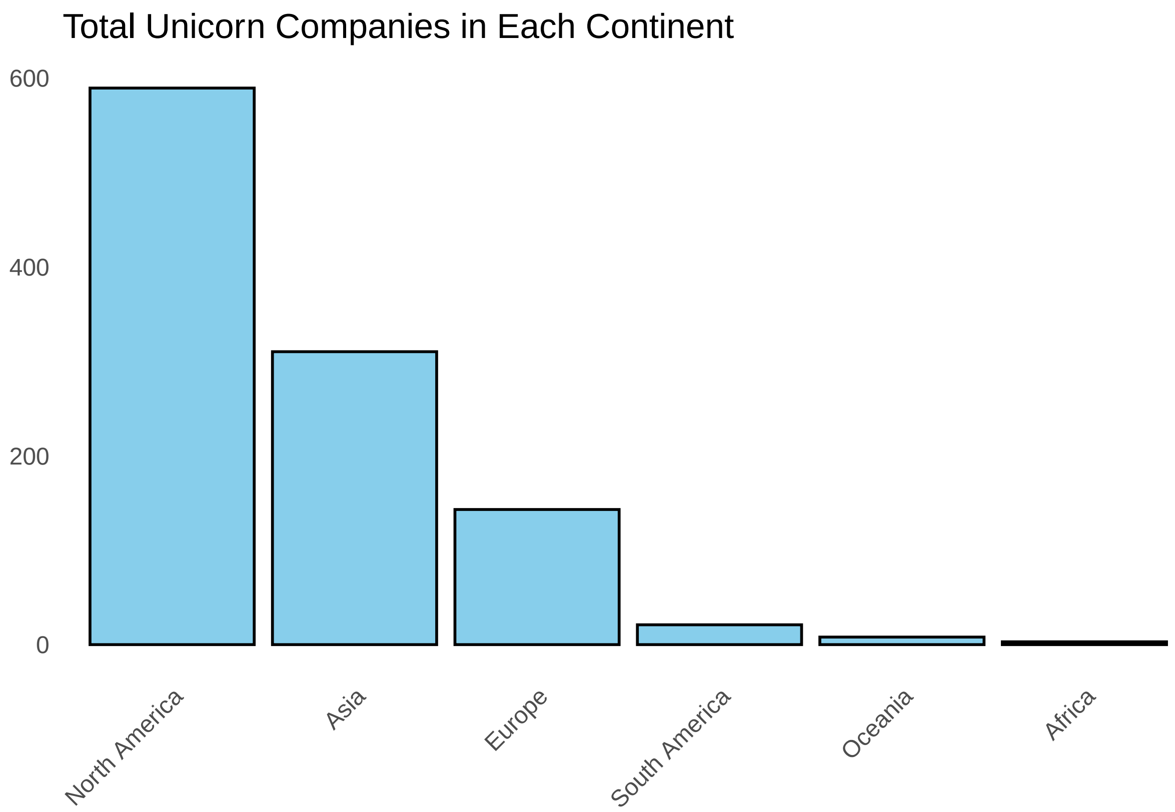
Exploratory Analysis

```
unicorn_summary <- unicorn %>%
  group_by(industry) %>%
  summarise(total_ind= n()) %>%
  arrange(-total_ind)

# create a bar chart
ggplot(unicorn_summary, aes(x = reorder(industry, -total_ind), y = total_ind)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_label(aes(label = total_ind), vjust = 1.0) +  # Provide labels using aes()
  labs(title = "Unicorns in Each Industry",
       x = "Industry",
       y = "Number of companies") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.grid = element_blank(), axis.text.y =
```



The Fintech industry which has experienced a massive boom in recent years unsurprisingly has the highest number of unicorn companies with 224, closely followed by the Internet software & services industry with 205 unicorn companies. This represents a significant lead over other startup companies in the unicorn space. When compared to traditional industries like Consumer & retail, travel, etc, the difference in the number of unicorn companies becomes apparent, highlighting a clear preference for investment and growth in these booming sectors.

```r
country_sum <-  unicorn %>%
  group_by(continent) %>%
  summarise(total_com = n()) %>%
  arrange(-total_com) %>%
  head(10)
ggplot(country_sum, aes(x = reorder(continent, -total_com), y = total_com)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Total Unicorn Companies in Each Continent",
       x = "",
       y = "") +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.grid = element_blank
```



The bar chart illustrates the distribution of unicorn companies across continents, revealing interesting insights into their prevalence in different regions.

Unsurprisingly, North America leads the way with the most unicorn companies (589), with majority of these companies domiciled in the United States. The vibrant tech ecosystem in North America particularly in regions like Silicon Valley, contributes significantly to this dominance.

Following North America, Asia and Europe exhibit substantial numbers of unicorn companies. Asia, driven by technology hubs in countries like China and India, secures the second position, closely followed by Europe.

In contrast, Africa is home to the fewest unicorn companies among the continents analyzed. While the tech landscape is growing in Africa, it currently faces challenges that impact the number of unicorn startups compared to other continents.

To analyze this dataset further, each company has separate investors and this column needs to be cleaned to ensure accuracy of analysis

```r
# replace every instance of double quotes with an empty string
unicorn <- unicorn %>%
  mutate(investor =
gsub("\"", "", unicorn$select_investors)) %>%
  select(-select_investors)
# separate each investor to a different row
unicorn_clean <- unicorn %>%
  mutate(investors = str_split(investor, ",")) %>%
  unnest() %>%
  select(-investor)
```

Valuation Analysis

In this section, I analyze the valuation column in the dataset by calculating different descriptive statistic measures, aggregating by different variables, to highlight potential outliers and understand the dataset better.

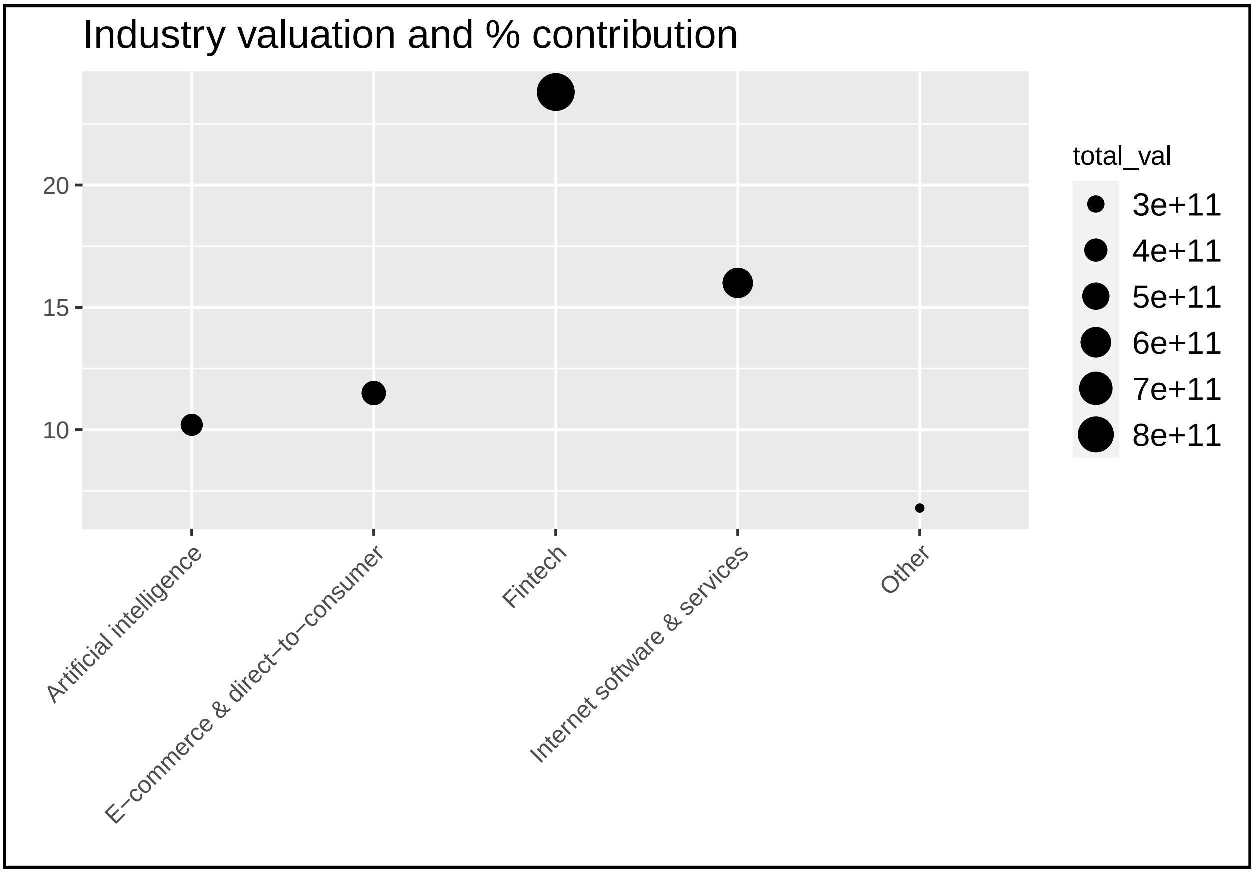
Summary

- The overall average Unicorn company valuation is about **$3.45B**
- Valuation by country:
  - The best performer here is **Bahamas**. It has one unicorn company, **FTX** (now defunct) – a cryptocurrency exchange and crypto hedge fund, which is valued at **$32B**.
  - **Sweden** is the next in line with an average valuation of **$10.5B** across its 6 unicorn companies. These are the only 2 countries that crossed the $10B mark in terms of average valuation.
  - **Croatia, Italy, and Czech Republic** all come last on this list with an average unicorn valuation of **$1B**.
- Valuation across Continents:
  - North America, with **$2.03 Trillion** contributes ~**54.8%** of total valuation
  - Asia, with **$1.07 Trillion** contributes ~**28.8%**
  - Europe, with **$503 Billion** contributes ~**13.6%**
  - Oceania, with **$56 Billion** contributes ~**1.5%**
  - South America, with **$48 Billion** contributes ~ **1.3%**
  - Africa, with **$5 Billion** contributes ~**0.1%**

```r
# average market valuation
unicorn %>%
  summarise(avg_valuation = mean(valuation))
```

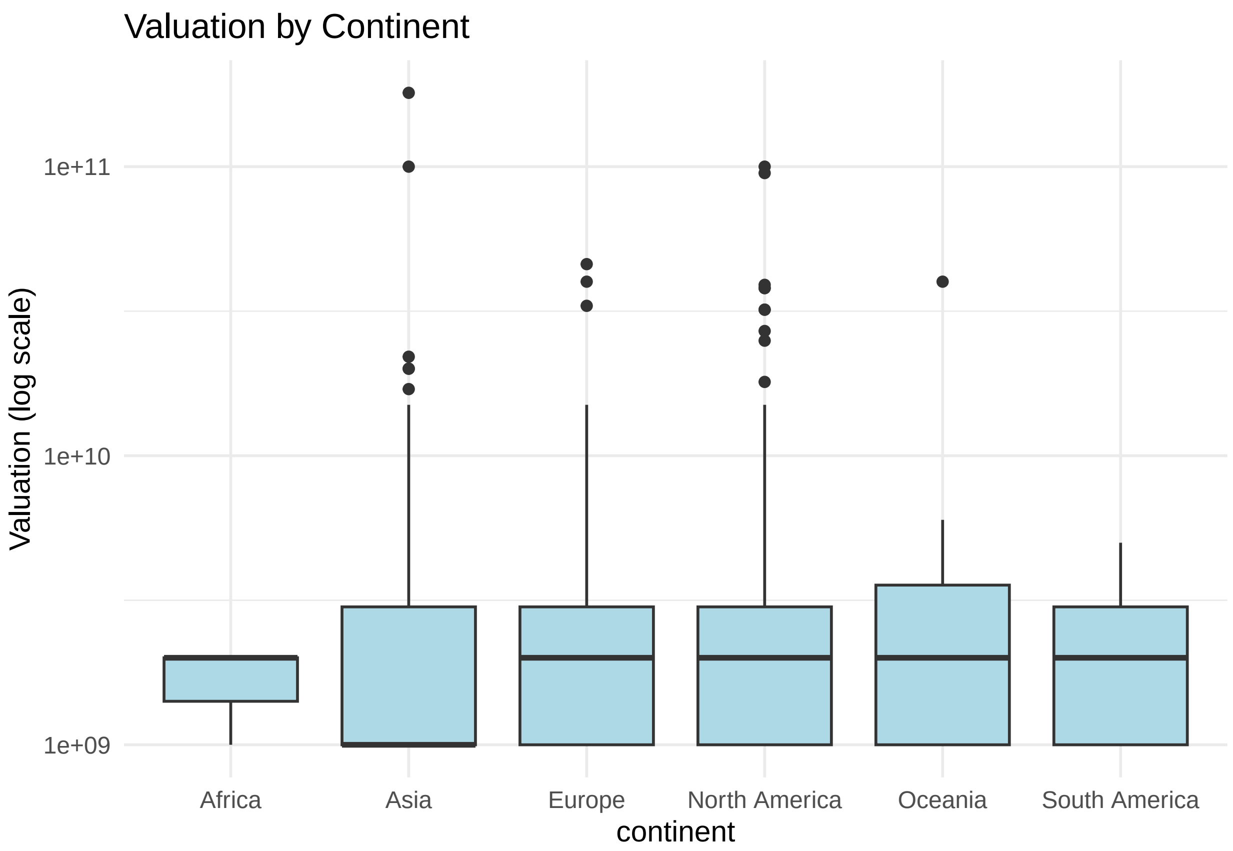
```
## # A tibble: 1 x 1
##   avg_valuation
##           <dbl>
## 1   3455307263.
```

```r
## average valuation for companies in each country
av_countr <- unicorn %>%
  group_by(country) %>%
  summarise(avg_country_valuation = mean(valuation)) %>%
  arrange(-avg_country_valuation)

## estimating each continent's contribution to total valuation
continent_val <- unicorn %>%
  group_by(continent) %>%
  summarise(total_val = sum(valuation), pct_contribution = round(total_val/sum(unicorn$valuation)*100,1
  arrange(-pct_contribution)

## estimate each industry's valuation contribution
unicorn %>%
  group_by(industry) %>%
  summarise(total_val = sum(valuation), pct_contribution = round(total_val/sum(unicorn$valuation)*100,1
  arrange(-pct_contribution) %>%
  head(5) %>%  # plot a bubble chart
  ggplot(aes(industry, y = pct_contribution, size = total_val)) + geom_point() + theme(axis.text.x = ele
    plot.background = element_rect(size = 1, color = "black"),
    panel.background = element_rect(fill = "#EAEAEA"),
    plot.title = element_text(size = 15),
    axis.title = element_text(size = 7),
    axis.text = element_text(size = 9),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 12)
  )
```



As expected, the **Fintech** bubble is the biggest, indicating its significant contribution to the valuation landscape.Further exploration of investment opportunities in this sector could be beneficial. **Internet software & services**, a relative oldhead in the unicorn space still looks in good shape, and you can hardly go wrong by investing into this space.

```
ggplot(unicorn, aes(continent, valuation)) + geom_boxplot(fill = "lightblue") + scale_y_log10() +theme_
```

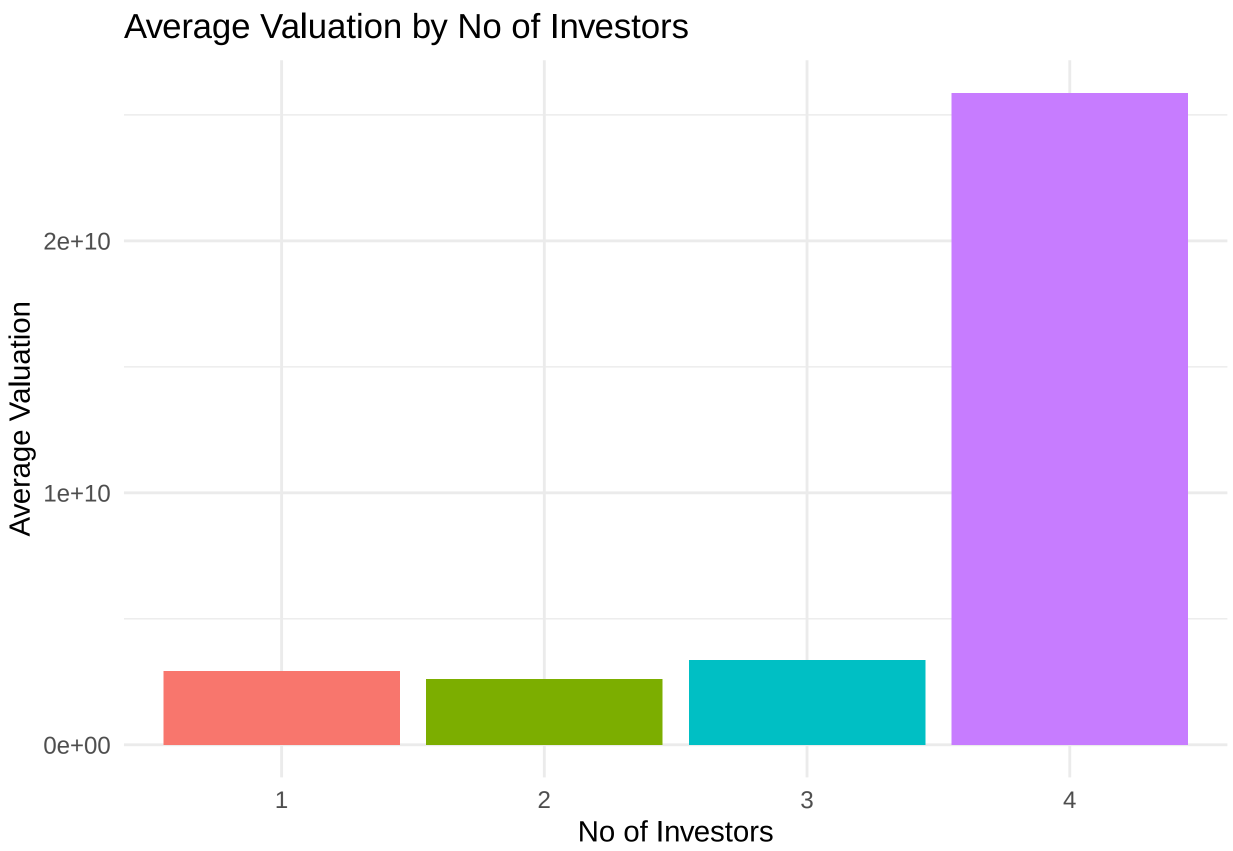


Investor Analysis

In this section, I want to understand if the valuation of a unicorn company is greatly affected by the number of investors the company has, and what the correlation is. Also, how the presence of some specific investors influences a company's valuation.

```
unicorn_clean %>%
  group_by(company, industry) %>%
  mutate(count = n()) %>%
  ungroup() %>%
  group_by(count) %>%
  summarise(average_valuation = mean(valuation)) %>%
  arrange(desc(average_valuation)) %>%

  ggplot(aes(x = factor(count), y = average_valuation, fill = factor(count))) +
  geom_bar(stat = "identity") +
  scale_x_discrete(name = "No of Investors") +
  scale_y_continuous(name = "Average Valuation") +
  ggtitle("Average Valuation by No of Investors") +
  theme_minimal() + guides(fill = FALSE)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
```



Key Findings

Our analysis revealed that companies with 4 investors exhibited the highest average valuation among unicorn companies. This suggests a positive correlation between the number of investors and the overall valuation of a company. A more in-depth examination of these companies and their unique characteristics could provide insights into the factors driving such high valuations.

Following closely behind, companies with 3 investors displayed the second-highest average valuation. This indicates a substantial valuation impact even with a slightly lower number of investors.

Interestingly, companies with only 1 investor secured the third-highest average valuation. This counter-intuitive finding suggests that individual investors, under certain circumstances, may contribute significantly to a company's valuation.

On the other hand, companies with 2 investors exhibited the lowest average valuation among the groups we analyzed. This prompts further investigation into potential reasons behind this lower valuation and whether it is influenced by specific industry dynamics or company characteristics.

```
inv <- unicorn_clean %>%
  group_by(company, industry) %>%
  mutate(count = n()) %>%
  ungroup() %>%
  group_by(count) %>%
  summarise(average_valuation = mean(valuation)) %>%
  arrange(desc(average_valuation))

cor(inv$count, inv$average_valuation)
```

```
## [1] 0.7839556
```

With a correlation of 0.78, this suggests a strong positive correlation between the number of investors and the expected average valuation of the company. It is important to note that correlation =/= causation.

```
# remove leading and trailing spaces in the investors column
unicorn_clean$investors <-  stringr::str_trim(unicorn_clean$investors)

#
unicorn_clean %>%
  group_by(investors) %>%
  summarise(total = n()) %>%
  arrange(-total) %>%
  head(5)
```

```
## # A tibble: 5 x 2
##   investors                total
##   <chr>                    <int>
## 1 Accel                       60
## 2 Andreessen Horowitz         53
## 3 Tiger Global Management     53
## 4 Sequoia Capital China       48
## 5 Insight Partners            47
```

The top 3 biggest investors are: - Accel (60) - Andreeessen Horowitz (53) - Tiger Global Management (53)

Further analysis can be conducted to understand the nature of investments of the heavy hitters and how profitable they are.

Conclusion

In summary, our analysis of the unicorn companies dataset has unearthed key insights into the dynamics of this unique startup ecosystem:

1. **Industry Dynamics**:

- Fintech leads the unicorn landscape, indicating a pronounced trend towards investment in innovative and high-growth sectors.
- Traditional industries exhibit fewer unicorn companies, emphasizing the shift in focus towards technology-driven ventures.

2. **Geographic Trends**:

- North America dominates with a significant number of unicorn companies, driven by robust tech ecosystems. Asia and Europe follow closely, reflecting the global distribution of high-valuation startups. Africa, while growing, currently lags behind in the number of unicorn companies.

3. **Valuation Analysis**:

- The average unicorn company valuation is around $3.45 billion. Valuations vary by country, with the Bahamas and Sweden leading the pack.

4. **Investor Influence**:

- A positive correlation (0.78) between the number of investors and average valuation suggests investor impact on a company's success.
- Companies with 4 investors show the highest average valuation, while those with 2 investors exhibit the lowest.

Recommendations:

- **Industry Focus**: Given the significant contribution of Fintech to unicorn valuations, consider exploring investment opportunities in this thriving sector.
- **Geographical Considerations**: Prioritize regions with strong tech ecosystems, such as North America and Asia, for potential unicorn investments.
- **Investor Insights**: Study the strategies of top investors like Accel, Andreessen Horowitz, and Tiger Global Management for potential investment guidance.

These recommendations align with the initial scenario of predicting a company's potential to reach a valuation over $5 billion based on characteristics such as its origin, industry, and investor details. By focusing on emerging sectors, strategic geographical locations, and understanding the impact of investor involvement, your company can enhance its ability to make informed and lucrative investment decisions in the dynamic unicorn startup landscape.

Project by: Ajanaku Ayomide image attribute: Image by rawpixel.com on Freepik