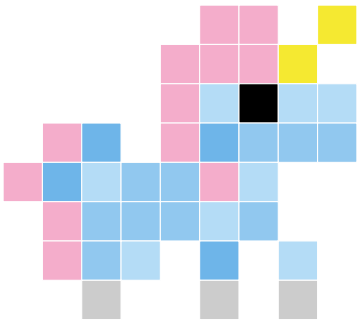

Exploring Unicorn Companies

Data Science · Mr. Merrick · January 15, 2026

For this assignment we will use `unicorn.csv` . A unicorn company is a private company with an estimated valuation of at least $1 billion. This dataset was pulled from Kaggle: Unicorn Companies Dataset.

Key variables

- `Company`
- `Valuation`
- `Country`, `City`
- `Industry`
- `Select_investors`
- `Founded_year`

Guiding questions

Answer each question using R. Include plots and a few sentences of interpretation.

Q1. Valuations: distribution.
Visualize and describe the distribution of unicorn valuations.

Q2. Top 5 countries by unicorn count.
Which five countries have the largest number of unicorn companies?

Q3. Valuations across the top 5 countries.
Using the five countries from Q2, compare valuations using side-by-side boxplots.

Q4. Top 5 industries by unicorn count.
Which five industries have the largest number of unicorn companies? Which industry would you most like to work in, and why?

Q5. Valuations across the top 5 industries.
Using the five industries from Q4, compare valuations using side-by-side boxplots.

Q6. Highest-valued unicorns.
List the 5 (or 10) companies with the highest valuations. Which one would you most like to work for, and why?

Q7. Change over time.
Has the number of unicorn companies (based on `Founded_year`) changed over time? Create a bar-chart showing the number of companies founded over the years.

Q8. Reflection.
If you sold a company for $1 billion, what would you spend the money on? Write a quick blurb and email it to your teacher.

Tutorial solutions (R)

```r
################################################################
# Exploring Unicorn Companies - R Solution Script
# Data file: unicorn.csv
################################################################

# ------------------------------
# Setup: packages + data
# ------------------------------
library(tidyverse)
library(scales)

uni <- read.csv("unicorn.csv")

# Quick look
glimpse(uni)


################################################################
# Q1) Visualize and describe the distribution of valuations
################################################################

# Histogram
uni %>%
  ggplot(aes(x = Valuation)) +
  geom_histogram(bins = 30) +
  labs(
    title = "Distribution of Unicorn Valuations",
    x = "Valuation (Billions of USD)",
    y = "Number of Companies"
  ) +
  theme_classic()

################################################################
# Q2) Which five countries have the largest number of unicorns?
################################################################

top_countries <- uni %>%
  group_by(Country) %>%
  summarize(counts = n()) %>%
  arrange(desc(counts))

head(top_countries, 5)

################################################################
# Q3) Do the top five countries differ in valuations on average?
# Use side-by-side boxplots.
################################################################

# Filter the dataset to only include the top 5 countries
uni_top_countries <- uni %>%
  filter(Country %in% top_countries$Country)

# Side-by-side boxplots of valuations by country
ggplot(uni_top_countries, aes(x = Country, y = Valuation, fill = Country)) +
```

```r
  geom_boxplot() +
  labs(
    title = "Unicorn Valuations in the Top 5 Countries",
    x = "Country",
    y = "Valuation (Billions of USD)"
  ) +
  theme_classic() +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1),
    legend.position = "none"
  )

# Numerical summaries (mean and median) by country
uni_top_countries %>%
  group_by(Country) %>%
  summarise(
    n = n(),
    mean_valuation = mean(Valuation, na.rm = TRUE),
    median_valuation = median(Valuation, na.rm = TRUE)
  ) %>%
  arrange(desc(mean_valuation))


################################################################
# Q4) Which five industries have the most unicorns?
################################################################

top_industries <- uni %>%
  count(Industry, sort = TRUE) %>%
  slice_head(n = 5)

top_industries

################################################################
# Q5) Do the top five industries differ in valuations on average?
# Use side-by-side boxplots.
################################################################

# Filter the dataset to only include the top 5 industries
uni_top_industries <- uni %>%
  filter(Industry %in% top_industries$Industry)

# Side-by-side boxplots of valuations by industry
ggplot(uni_top_industries, aes(x = Industry, y = Valuation, fill = Industry)) +
  geom_boxplot() +
  labs(
    title = "Unicorn Valuations in the Top 5 Industries",
    x = "Industry",
    y = "Valuation (Billions of USD)"
  ) +
  theme_classic() +
  theme(
    axis.text.x = element_text(angle = 35, hjust = 1),
    legend.position = "none"
  )

```

```r
# Numerical summaries (mean and median) by industry
uni_top_industries %>%
  group_by(Industry) %>%
  summarise(
    n = n(),
    mean_valuation = mean(Valuation, na.rm = TRUE),
    median_valuation = median(Valuation, na.rm = TRUE)
  ) %>%
  arrange(desc(mean_valuation))


################################################################
# Q6) Which 5 companies have the highest valuations?
################################################################

uni %>%
  arrange(desc(Valuation)) %>%
  select(Company, Valuation, Country, Industry, City) %>%
  slice_head(n = 10)


################################################################
# Q7) Has the number of unicorn companies changed over time?
# (Using Founded year) - Barchart-style
################################################################

uni %>%
  filter(!is.na(Founded_year)) %>%
  ggplot(aes(x = Founded_year)) +
  geom_bar() +
  labs(
    title = "Number of Unicorn Companies by Founded Year",
    x = "Founded Year",
    y = "Number of Companies"
  ) +
  theme_classic()

################################################################
# Q8) Reflection question
# No code needed (written response).
################################################################
```