Decoding Success: An In-depth Analysis of the 2019 INC 5000 Companies

Code

Tanvir Pahwa 100889937

Introduction

The dataset I chose to analyze on can be found in the `INC 5000 Companies 2019.csv` file. I saw this dataset on Kaggle and I instantly knew this was the dataset that I wanted to work on as I have been very interested in finance over the past month. The INC 5000 list is a list of the fastest-growing private companies in America, showing some very important data such as their growth rate, their revenue and their location!

The questions we are aiming to answer are:

1. Does the distribution of companies per state make sense? Are their too many companies in one place? What factor brings them there?
2. What industry would be the best to invest into?

We will be using these graphics to answer these questions and paint a clearer picture.

Data Manipulation

To start with I loaded all of the necessary libraries. I then cleaned the dataset by filtering out incomplete rows and excluding the data for Alaska and Hawaii due to their minimal impact. I also changed the revenue from word to standard notation for easier graphing, and grouped over 25 industries into 11 categories for improved readability in the graphics. I also added another column that shows the full state name as it is required for the map.

Hide

```
# Load necessary libraries
  library(tidyverse)
  library(maps)
  library(scales)
  library(ggthemes)
  library(tidycensus)


# Load the data
data <- read.csv("INC 5000 companies 2019.csv")
tidycensus::census_api_key("1e959b9a6de23896a02e0b2fdf2cd87bc132bffa", install = TRUE, overwrite = TRUE)
```

```
[1] "1e959b9a6de23896a02e0b2fdf2cd87bc132bffa"
```

Hide

```
data$state_full <- setNames(state.name, state.abb)[toupper(data$state)]

data <- data %>%
  na.omit() %>%
  filter(!(state_full %in% c("Alaska", "Hawaii"))) %>%
  mutate(revenue = sapply(gsub(" Million", "e6", gsub(" Billion", "e9", gsub(",", "", revenue))), as.numeri
c)) %>%
  mutate(industry_category = case_when(
    industry %in% c("Computer Hardware", "IT Management", "IT Services", "IT System Development", "Softwar
e", "Telecommunications") ~ "Technology",
    industry %in% c("Advertising & Marketing", "Business Products & Services", "Human Resources") ~ "Busine
ss Services",
    industry %in% c("Construction", "Engineering") ~ "Construction & Engineering",
    industry %in% c("Consumer Products & Services", "Food & Beverage", "Retail") ~ "Consumer Goods & Servic
es",
    industry %in% c("Education", "Health") ~ "Education & Health",
    industry %in% c("Energy", "Environmental Services") ~ "Energy & Environment",
    industry %in% c("Financial Services", "Insurance") ~ "Finance & Insurance",
    industry %in% c("Government Services", "Security") ~ "Government & Security",
    industry %in% c("Logistics & Transportation", "Manufacturing") ~ "Logistics, Transportation & Manufactu
ring",
    industry %in% c("Media", "Travel & Hospitality") ~ "Media & Hospitality",
    industry == "Real Estate" ~ "Real Estate"
  ))
```

Creating the map

The map shows us how many companies there are in each of the Contiguous United States.

Hide

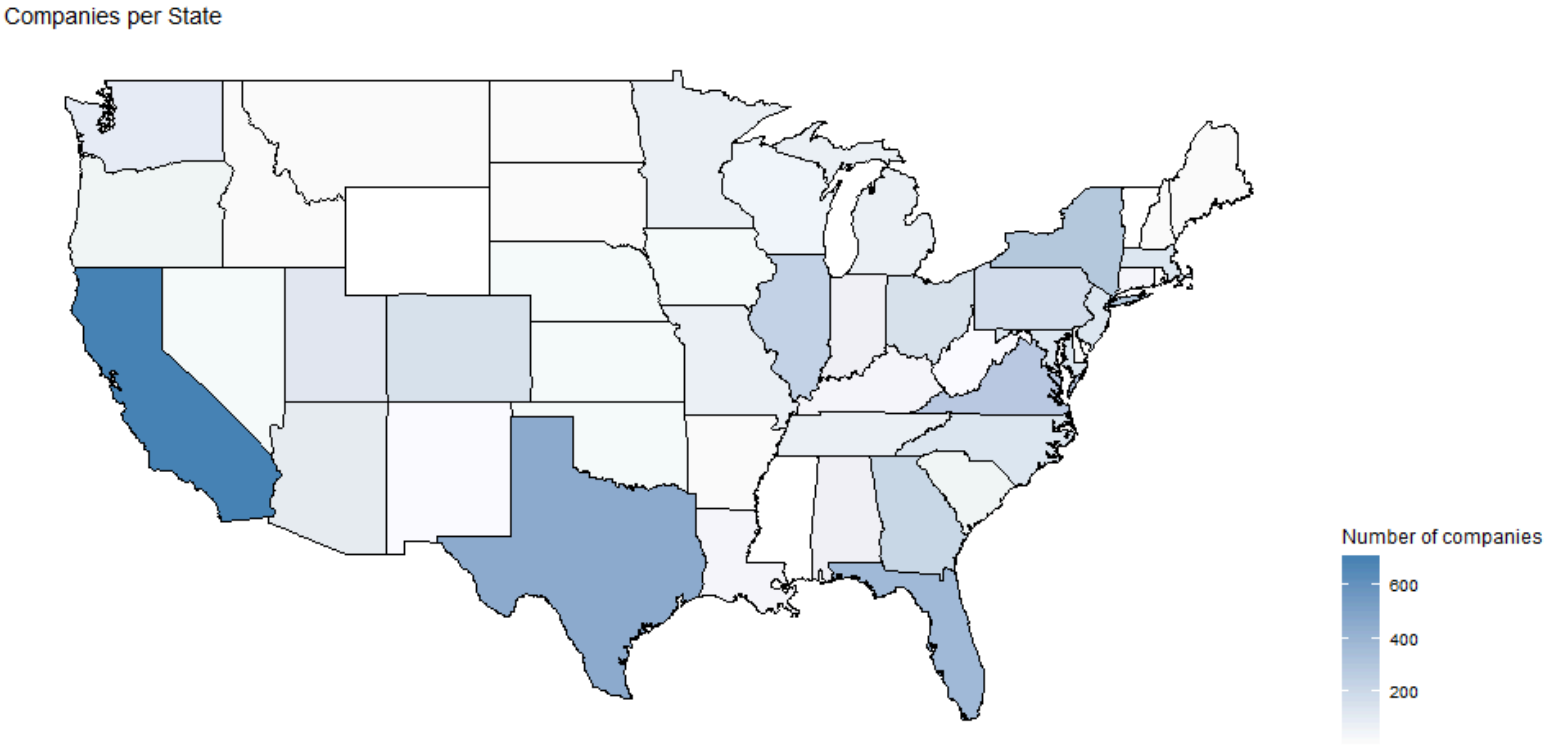
```
# Gathering data to be used in the map
state_counts <- data %>%
  group_by(state_full) %>%
  summarise(n = n())

states <- map_data("state")

# Convert the state names in both datasets to lower case
states$region <- tolower(states$region)
state_counts$state_full <- tolower(state_counts$state_full)

# Merge data with the states data
mapped_data <- left_join(states, state_counts, by = c("region" = "state_full"))

# Create a choropleth map of the contiguous USA
ggplot() +
  geom_polygon(data = mapped_data, aes(x=long, y = lat, group = group, fill = n), color="black") +
  scale_fill_gradient(low = "white", high = "steelblue") +
  coord_fixed(1.3) +
  theme_map() +
  theme(legend.position = "right") +
  labs(title = "Companies per State", fill = "Number of companies")
```



The companies per state map shows us that California has the most companies, followed closely by Texas, Florida and New York. This outcome is somewhat expected as all four of these states are the most populated states in the U.S.A. We know this after we look at the following map.

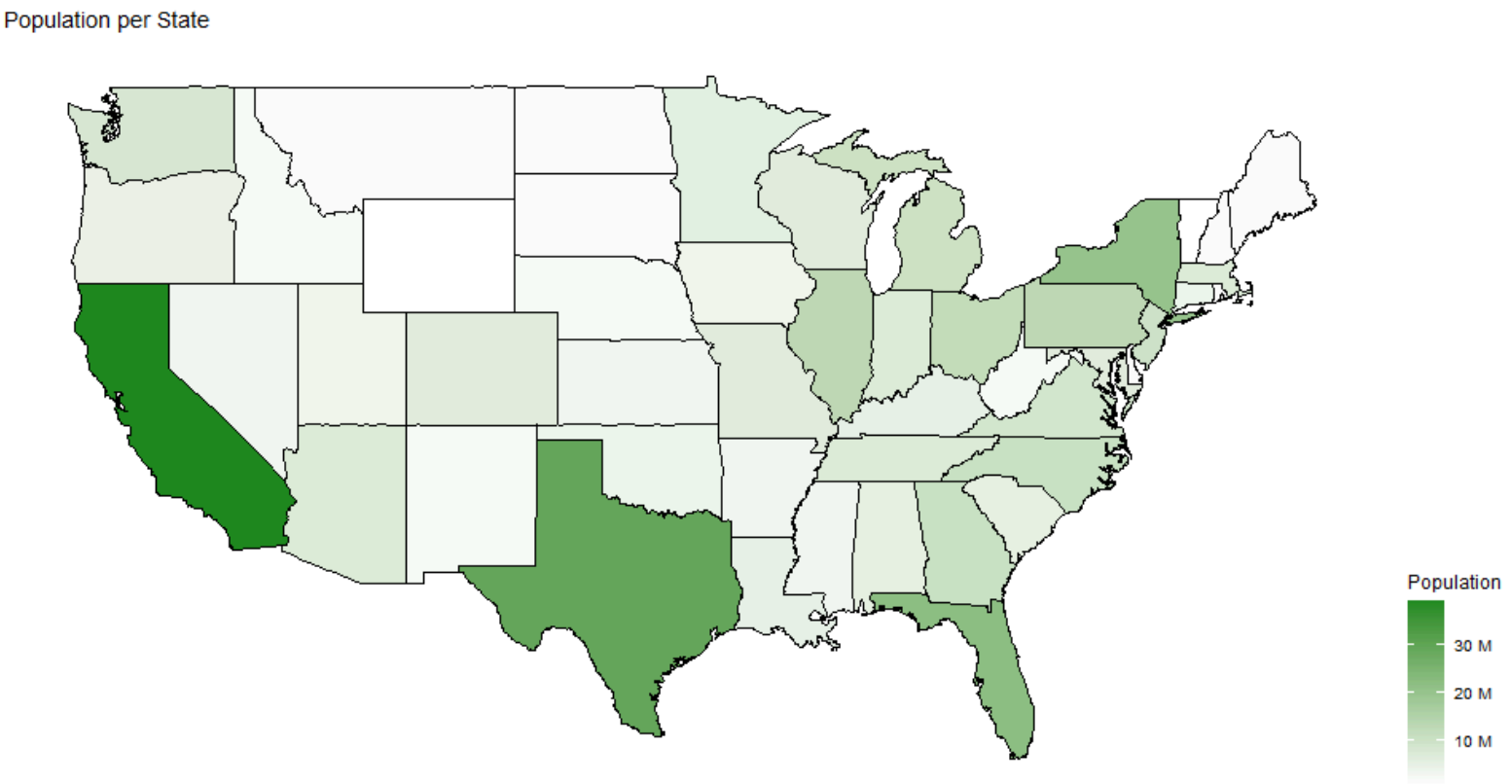
Hide

```
# Get the population data
population_data <- tidycensus::get_acs(geography = "state", variables = "B01003_001") %>%
  .[!(.$NAME %in% c("Alaska", "Hawaii", "Puerto Rico", "District of Columbia")), ]

# Convert the state names in population_data to lowercase
population_data$NAME <- tolower(population_data$NAME)

# Merge the population data with the existing mapped_data
merged_data <- left_join(mapped_data, population_data, by = c("region" = "NAME"))


# Create the choropleth map of the contiguous USA based on population
ggplot() +
  geom_polygon(data = merged_data, aes(x=long, y = lat, group = group, fill = estimate), color="black") +
  scale_fill_gradient(low = "white", high = "forestgreen", labels = scales::comma_format(scale = .000001, p
refix = "", suffix = " M")) +
  coord_fixed(1.3) +
  theme_map() +
  theme(legend.position = "right") +
  labs(title = "Population per State", fill = "Population")
```



This map is exclusively just a supporting map that shows the population of each of the contiguous states of America. The map was kept similar to the last graphic purely for ease of readability and comparability.

The distribution of companies actually makes a lot of sense. It does not seem to have too many companies in one place even though it might feel that way when you first see it. The factor that seems to “bring” companies to these states is likely the population.

Creating the pie chart

This pie chart shows the percentage of workers in the top 10 employed states.

Hide

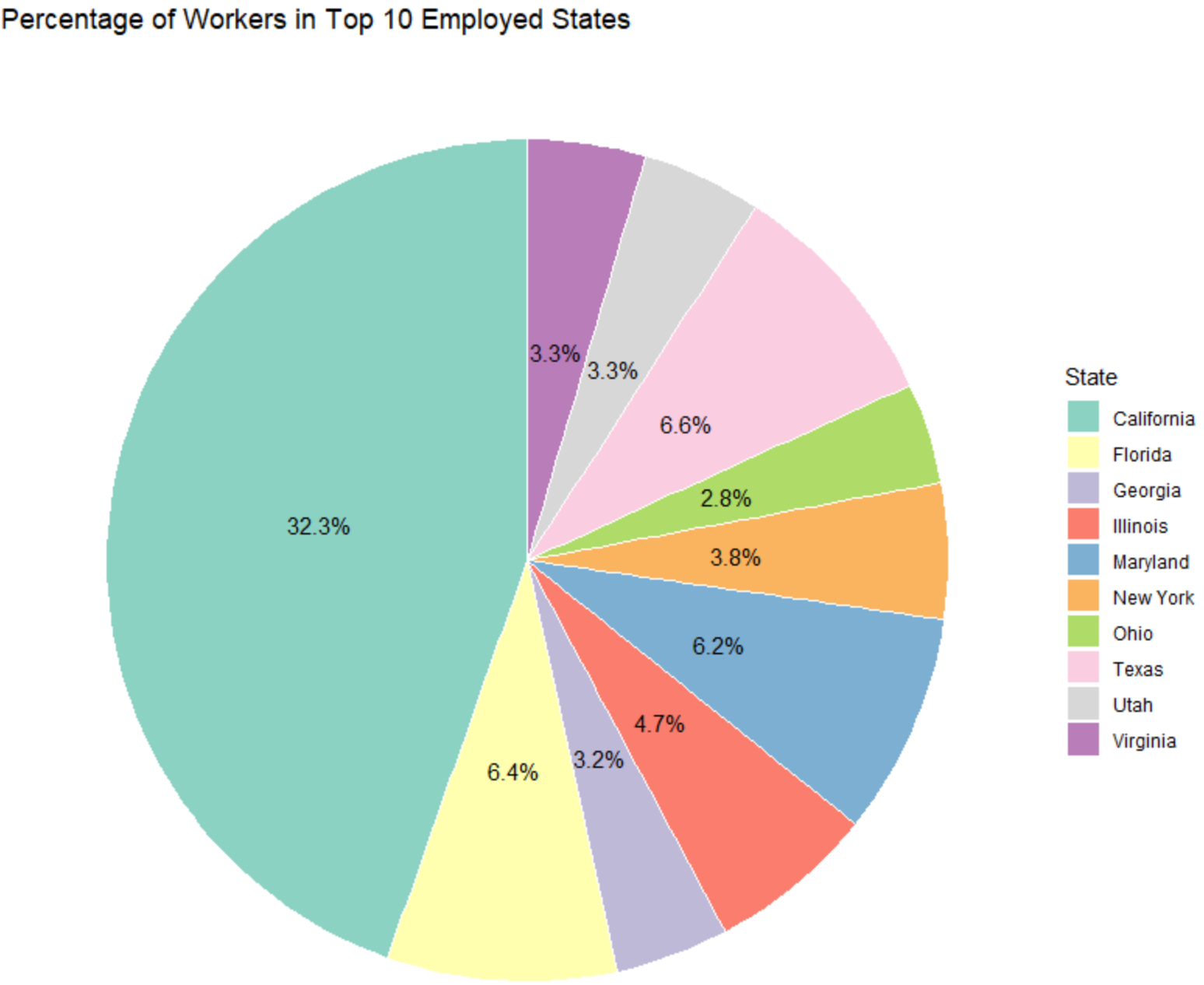
```
# Calculate the total number of workers per state
worker_counts <- data %>%
  group_by(state_full) %>%
  summarise(total_workers = sum(workers))

# Calculate the total number of workers in all states
total_workers_all <- sum(worker_counts$total_workers)

# Calculate the percentage of workers per state
worker_counts <- worker_counts %>%
  mutate(percentage = total_workers / total_workers_all * 100)

# Select the top 10 states with the most workers
top_states <- worker_counts %>%
  arrange(desc(total_workers)) %>%
  head(10)

# Create a pie chart
ggplot(top_states, aes(x = "", y = percentage, fill = state_full)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0) +
  theme_void() +
  theme(legend.position = "right") +
  scale_fill_brewer(palette = "Set3") +
  labs(fill = "State", title = "Percentage of Workers in Top 10 Employed States") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), position = position_stack(vjust = 0.5))
```



This shows even further that the distribution of companies and workers are not random, but are influenced by factors such as the population of a state.

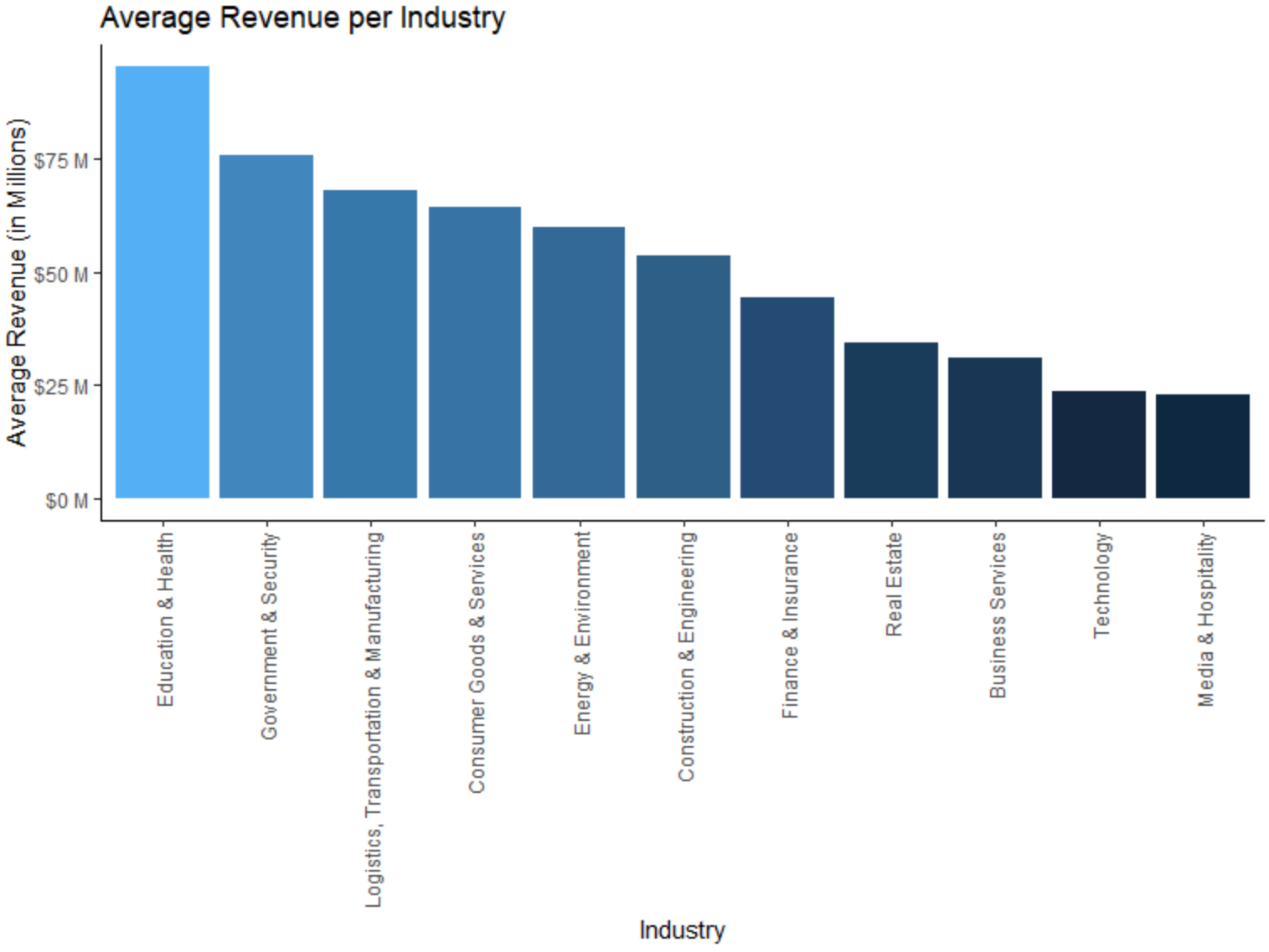
Creating the bar plot

This bar plot will show us the average revenue per industry.

Hide

```
# Calculate the averages by industry
avgs <- data %>%
  group_by(industry_category) %>%
  summarise(
    avg_growth = mean(growth_.),
    avg_revenue = mean(revenue)
  )

# Create a bar plot
ggplot(avgs, aes(x = reorder(industry_category, -avg_revenue), y = avg_revenue)) +
  geom_bar(stat = "identity", aes(fill = avg_revenue)) +
  scale_y_continuous(labels = scales::dollar_format(scale = .000001, prefix = "$", suffix = " M")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1), legend.position = "none") +
  labs(x = "Industry", y = "Average Revenue (in Millions)", title = "Average Revenue per Industry")
```



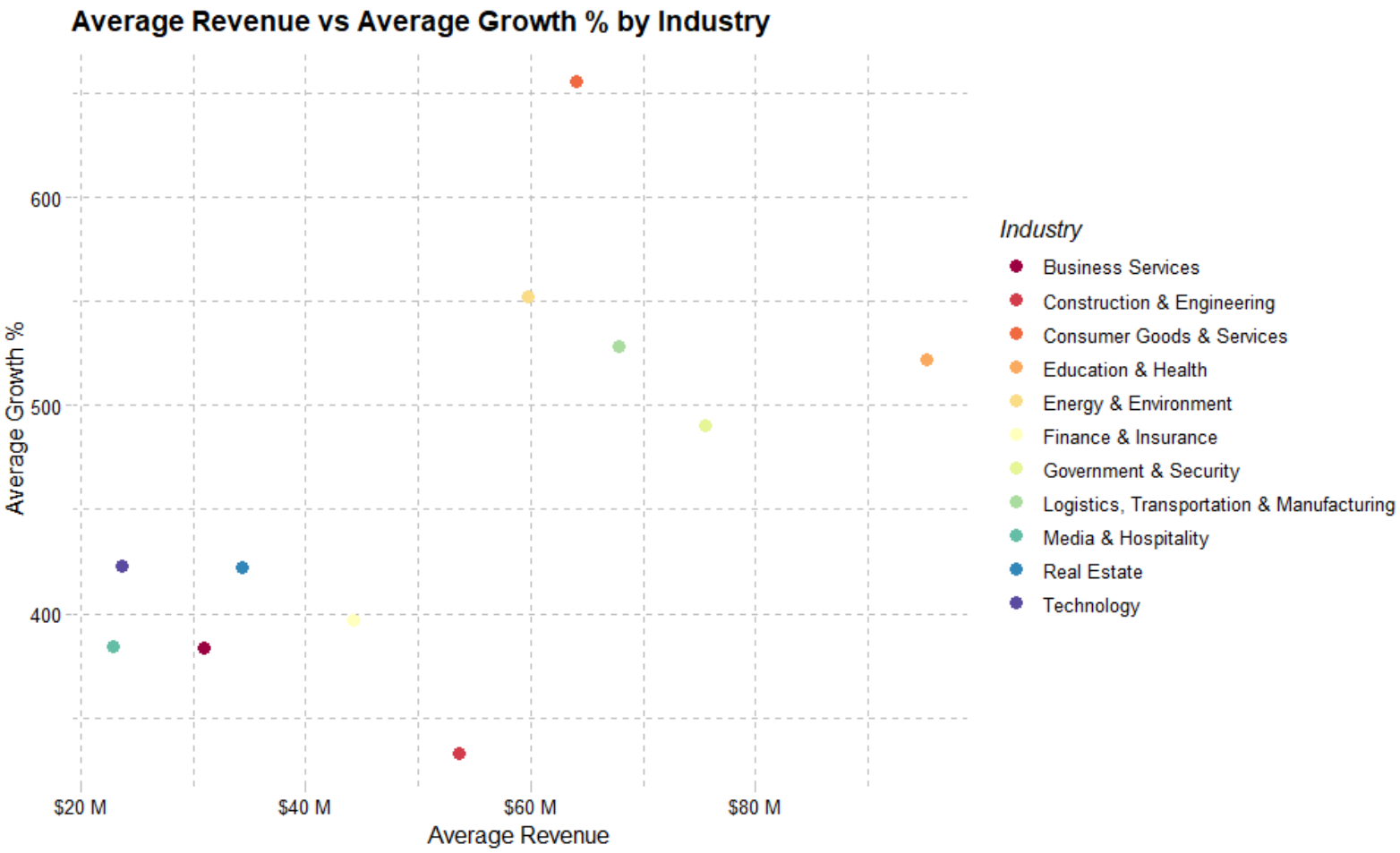
When looking at this graph you might think that Education & Health would be the best industry to invest into, this would not be entirely true as just revenue would not tell you the whole story.

Creating the scatter plot

This leads well into our next graphic. The scatter plot shows us the direct correlation between the average growth % and the average revenue for each industry.

Hide

```
# Create the scatter plot
ggplot(avgs, aes(x=avg_revenue, y=avg_growth, color=industry_category)) +
  geom_point(size=3) +
  scale_color_brewer(palette="Spectral") +
  scale_x_continuous(labels = scales::dollar_format(scale = .000001, prefix = "$", suffix = " M")) +
  labs(title="Average Revenue vs Average Growth % by Industry", x="Average Revenue", y="Average Growth %",
color="Industry") +
  theme_pander()
```



When investing you also have to think about how much money you would get out of it, that is where growth rate comes into play. The best industry to invest in according to this data would be Consumer Goods & Services. This is due to their very high average growth, and the relatively high average revenue. The next best industry would be Energy & Environment, followed closely by Logistics, Transportation & Manufacturing and Education & Health.

Conclusion

I feel that this analysis of the INC 5000 Companies 2019 dataset shows insightful trends. The distribution of companies across states follows the population density suggesting that the population level may be very relevant for company location. With respect to the industries of Consumer Goods & Services, Energy & Environment, Logistics, Transportation & Manufacturing, and Education & Health, they have shown positive and attractive investment potential due to their higher average growth and revenue. Nevertheless, both revenue and growth rate should be taken into account because they offer more complete information on an industry's performance. The above shows the relevance of data-driven decision-making in finance.

References

`INC 5000 Companies 2019.csv` (https://www.kaggle.com/datasets/mysarahmadbhat/inc-5000-companies/data)

`U.S. Population Data` (https://api.census.gov/data/key_signup.html)