Practice Scraping Data from Website

Tam Tran The

Find a table in Wikipedia that can be scraped and visualized. Interpret your graphical display.

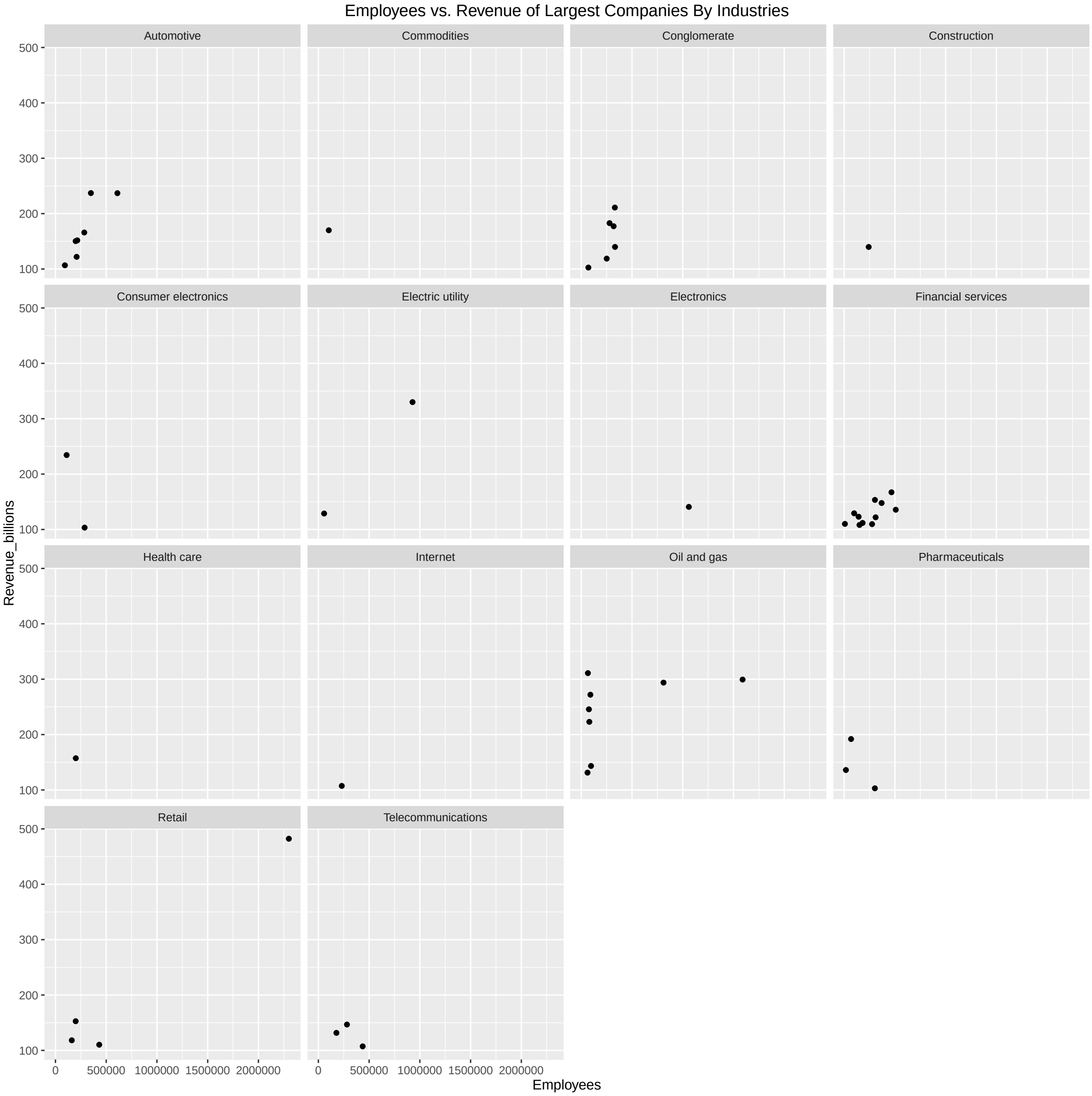
```r
# Import data from an html file
my_html <- read_html("https://en.wikipedia.org/wiki/List_of_largest_companies_by_revenue")
tables <- my_html %>% html_nodes(css = "table")
relevant_tables <- tables[grep("Walmart", tables)]
largest_companies <- html_table(relevant_tables[[1]], fill = TRUE)
head(largest_companies)
```

```
##   Ranking                     Name         Industry Revenue (USD billions)
## 1       1                  Walmart           Retail                   $482
## 2       2               State Grid Electric utility                   $330
## 3       3                  Samsung     Conglomerate                   $177
## 4       4             Saudi Aramco      Oil and gas                   $311
## 5       5 China National Petroleum      Oil and gas                   $299
## 6       6            Sinopec Group      Oil and gas                   $294
##   Revenue growth Employees Country          Headquarters            CEO
## 1           0.7% 2,300,000      NA Bentonville, Arkansas  Doug McMillon
## 2           2.9%   927,839      NA               Beijing    Shu Yinbiao
## 3           9.4%   319,000      NA                 Suwon   Oh-Hyun Kwon
## 4          40.1%    65,266      NA               Dhahran Amin H. Nasser
## 5          30.2% 1,589,508      NA               Beijing     Wang Yilin
## 6          34.1%   810,538      NA               Beijing      Wang Yupu
##   Ref(s)
## 1    [1]
## 2    [2]
## 3    [3]
## 4    [4]
## 5    [5]
## 6    [6]
```

```r
# Clean data
largest_companies <- largest_companies[, c(2:4,6)]
colnames(largest_companies)[3] <- "Revenue_billions"
largest_companies$Name = as.factor(largest_companies$Name)
largest_companies$Industry = as.factor(largest_companies$Industry)
largest_companies$Revenue_billions = as.numeric(gsub('\\$', '', largest_companies$Revenue_billions))
largest_companies$Employees = as.numeric(gsub(',', '', largest_companies$Employees))
head(largest_companies)
```

```
##                       Name         Industry Revenue_billions Employees
## 1                  Walmart           Retail              482   2300000
## 2               State Grid Electric utility              330    927839
## 3                  Samsung     Conglomerate              177    319000
## 4             Saudi Aramco      Oil and gas              311     65266
## 5 China National Petroleum      Oil and gas              299   1589508
## 6            Sinopec Group      Oil and gas              294    810538
```

```
# Visualize
ggplot(data = largest_companies, aes(x=Employees, y=Revenue_billions)) +
  geom_jitter() +
  facet_wrap(~Industry, ncol=4) +
  labs(title="Employees vs. Revenue of Largest Companies By Industries")
```



Interpretation: The industry which has the highest number of largest companies by revenue is financial services, followed by oil and gas, automotive and conglomerate. Most of these companies have fewer than 500,000 employees. Specifically, there are only 6 companies which have more than 500,000 employees. The relationship between number of employees and revenues is not clear from this scatterplot.