Forbes_Global_2000-2022

Pariya

In this project, I have cleaned the data to solve some important questions and made a visualized chart to know the trend of the data.

I have downloaded the data source from this link

https://data.world/aroissues/forbes-global-2000-2008-2019

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.3     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

Load the Excel file.

```
df <- read_excel("ForbesGlobal2000-2022.xlsx")
```

Check that this data set does not have na variable.

```
na_val <- is.na(df)
head(is.na(df)) #is.na(df) has many outputs, so I will show only a few of them.
```

```
##      Rank_nr Company Industry Country Sales Profits Assets Market_Value
## [1,]   FALSE   FALSE    FALSE   FALSE FALSE   FALSE  FALSE        FALSE
## [2,]   FALSE   FALSE    FALSE   FALSE FALSE   FALSE  FALSE        FALSE
## [3,]   FALSE   FALSE    FALSE   FALSE FALSE   FALSE  FALSE        FALSE
## [4,]   FALSE   FALSE    FALSE   FALSE FALSE   FALSE  FALSE        FALSE
## [5,]   FALSE   FALSE    FALSE   FALSE FALSE   FALSE  FALSE        FALSE
## [6,]   FALSE   FALSE    FALSE   FALSE FALSE   FALSE  FALSE        FALSE
```

```r
mean(na_val)
```

```
## [1] 0
```

0 mean do not have na value

Which country has the most Forbes companies?

```r
df_country <- df %>%
  group_by(Country) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
df_country$Country[1]
```

```
## [1] "United States"
```

The United States has the most Forbe companies.

Which five countries have the most market value?

```r
df_country_market <- df %>%
  select(Country, Market_Value) %>%
  group_by(Country) %>%
  summarize(total_market_value = sum(Market_Value)) %>%
  arrange(desc(total_market_value))
head(df_country_market, 5)
```

```
## # A tibble: 5 x 2
##   Country       total_market_value
##   <chr>                      <dbl>
## 1 United States           38185028
## 2 China                    6839165
## 3 Japan                    3440118
## 4 Canada                   2868302
## 5 Saudi Arabia             2809463
```

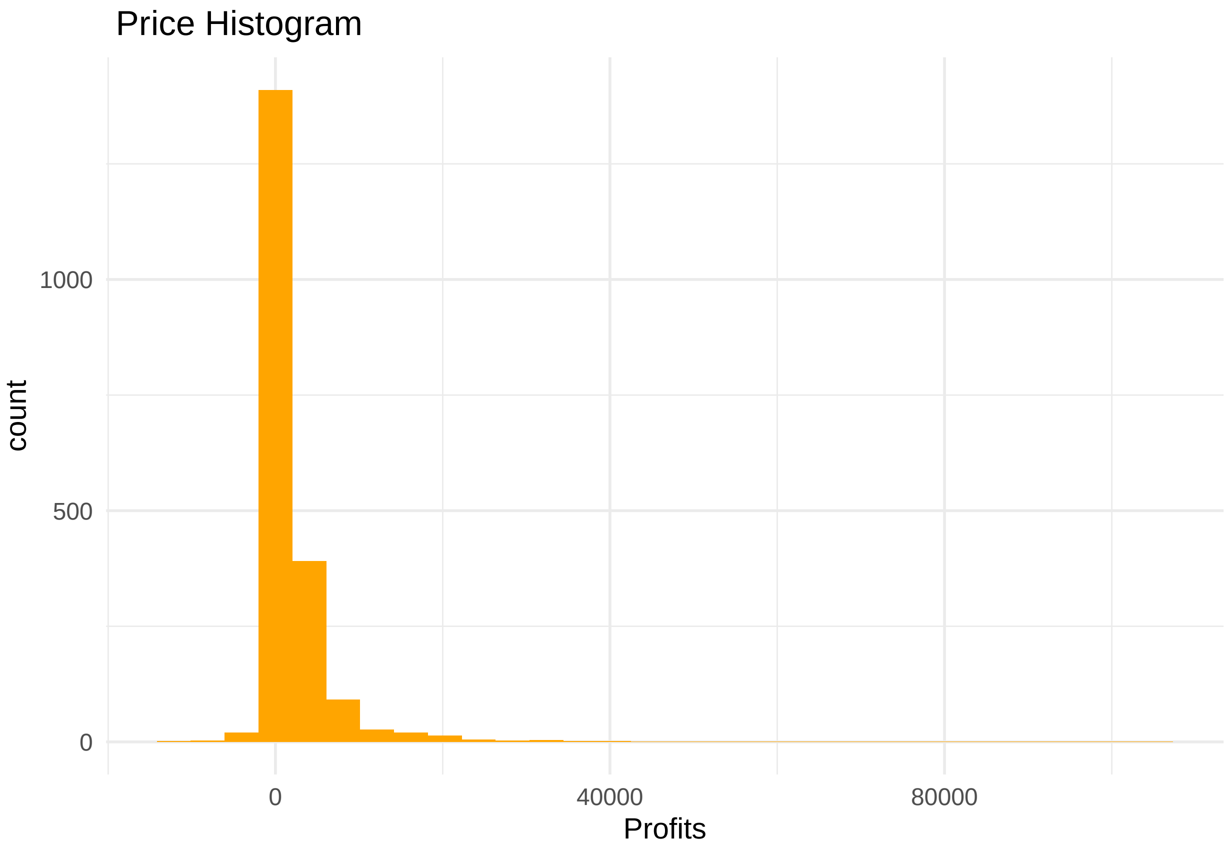
Which industry has the most companies?

```r
df_industry <- df %>%
  group_by(Industry) %>%
  summarize(total_num = sum(n= n()))
df_industry[1,1]
```

```
## # A tibble: 1 x 1
##   Industry
##   <chr>
## 1 Aerospace & Defense
```

Plot the histogram of prices

```
ggplot(df, aes(Profits)) +
  geom_histogram(bins = 30, fill ="orange")+
  theme_minimal()+
  labs(title = " Price Histogram",
       )
```

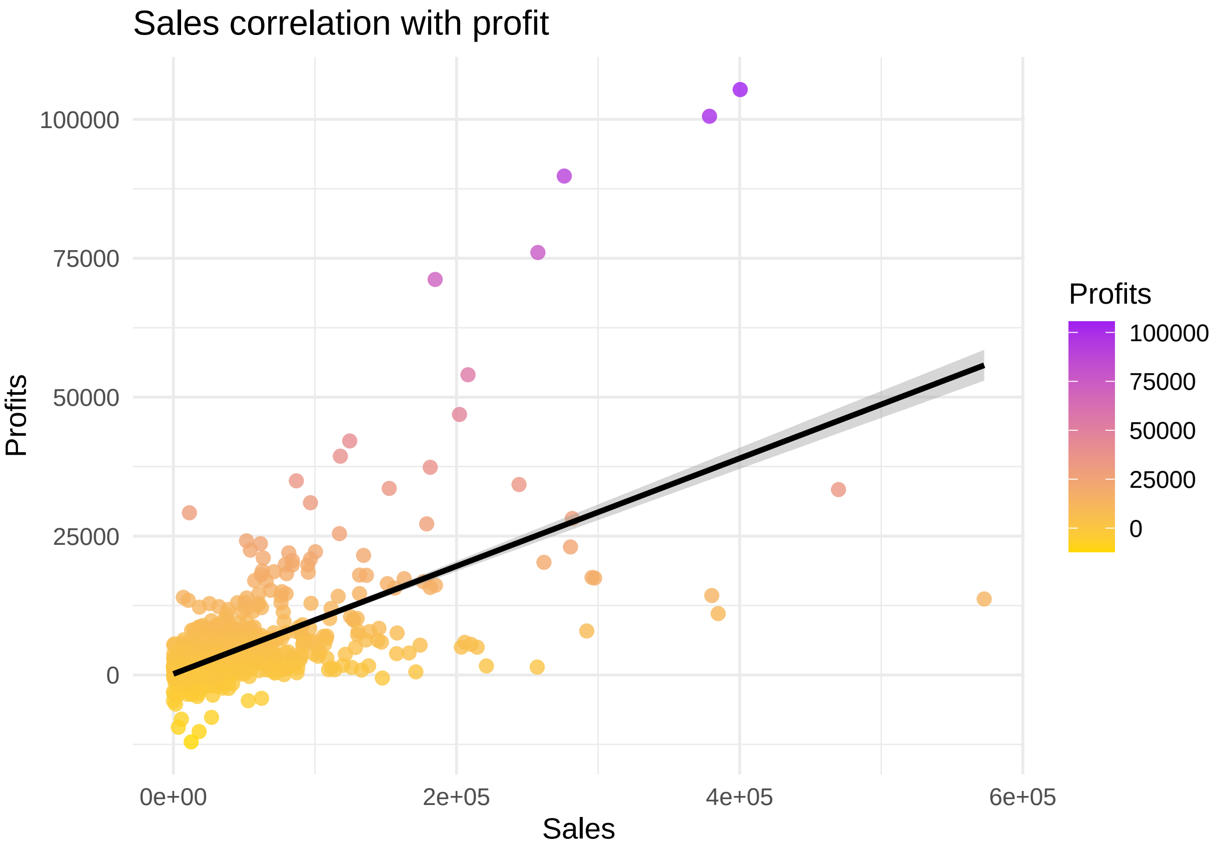


The distribution of prices is right-skewed.

Plot the sales correlation with profits.

```
ggplot(df, aes(x= Sales, y = Profits, col = Profits)) +
  geom_point(size = 2, alpha= 0.8) +
  scale_color_gradient(low = "gold", high = "purple") +
  labs(title = "Sales correlation with profit") +
  theme_minimal() +
  geom_smooth(method = "lm",
              col = "black")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



When sales increase, profits also increase.

Which industry has the best profits?

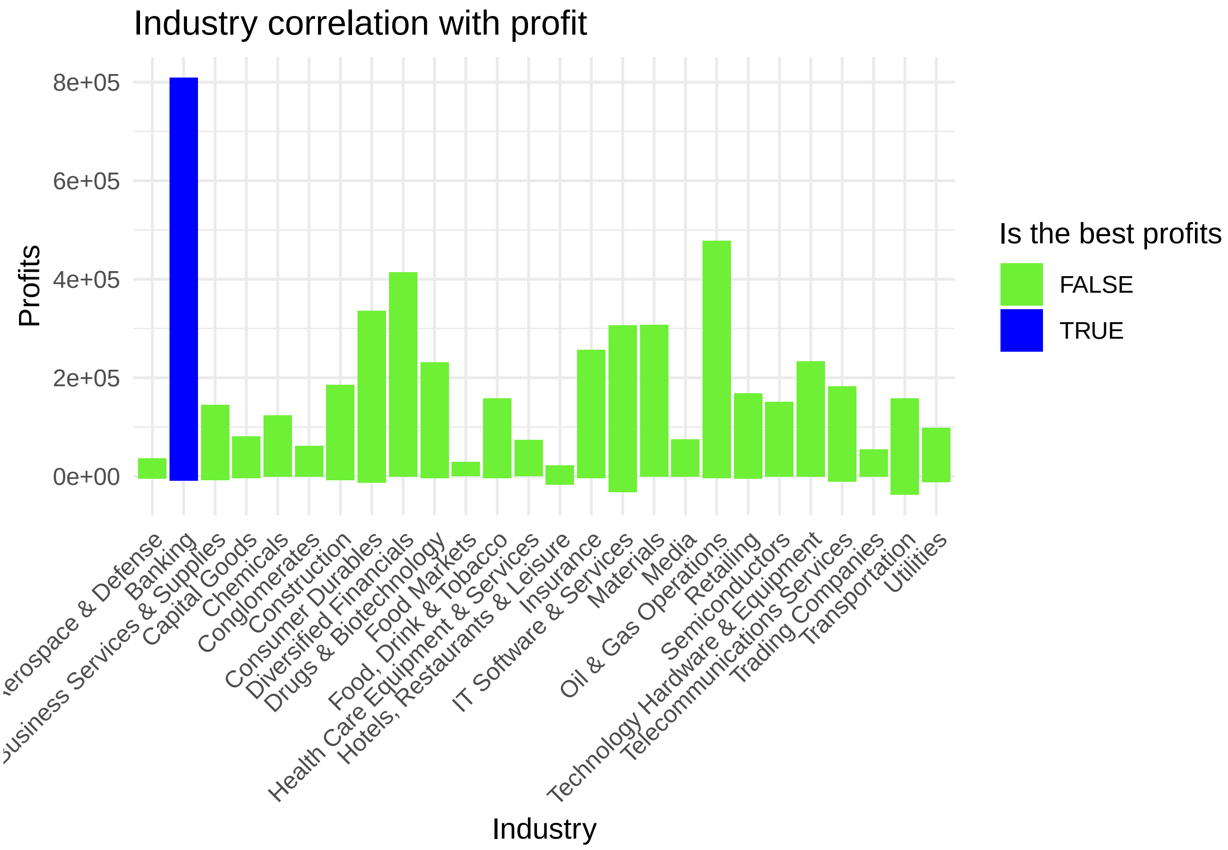
```
industry_group <- df %>%
  select(Industry, Profits) %>%
  group_by(Industry) %>%
  summarize(total_pro_indus = sum(Profits)) %>%
  arrange(desc(total_pro_indus))
industry_group[1,1]
```

```
## # A tibble: 1 x 1
##   Industry
##   <chr>
## 1 Banking
```

The best industry with the most profit is banking, so I will check the answer with a bar chart.

Plot the industry correlation with profits.

```
ggplot(df, aes(x = Industry, y = Profits, fill = Industry == industry_group$Industry[1])) +
  geom_col() +
  scale_fill_manual(values = c("TRUE" = "blue", "FALSE" = "#6EF036"),
                    name =  "Is the best profits") +
  theme_minimal() +
  labs(title = "Industry correlation with profit") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Industry correlation with profit
Profits
8e+05
6e+05
4e+05
2e+05
0e+00
erospace & Defense
Banking
usiness Services & Supplies
Capital Goods
Chemicals
Conglomerates
Construction
Consumer Durables
Diversified Financials
Drugs & Biotechnology
Food Markets
Food, Drink & Tobacco
Health Care Equipment & Services
Hotels, Restaurants & Leisure
Insurance
IT Software & Services
Materials
Media
Oil & Gas Operations
Retailing
Semiconductors
Technology Hardware & Equipment
Telecommunications Services
Trading Companies
Transportation
Utilities
Industry
Is the best profits
🟩 FALSE
🟦 TRUE