Final Project - I want to be a Billionaire

DATA LOADING, DATA STRUCTURE

```
'install.packages("tidyverse")
install.packages("countrycode")
install.packages("moderndive")'
```

```
## [1] "install.packages(\"tidyverse\")\ninstall.packages(\"countrycode\")\ninstall.packages(\"moderndive\")"
```

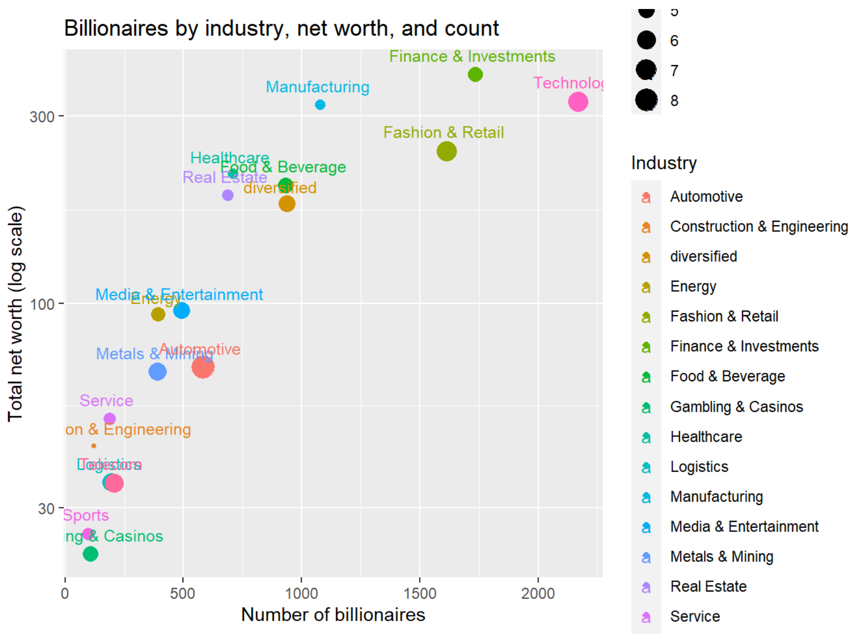
```
library(dplyr)
library(ggplot2)
library(countrycode)
library(moderndive)

baires <- read.csv("Forbes Billionaires.csv")
baires$continent <- countrycode(sourcevar = baires[, "Country"],
                            origin = "country.name",
                            destination = "continent")

str(baires)
```

```
## 'data.frame':    2600 obs. of  8 variables:
##  $ Rank     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name     : chr  "Elon Musk " "Jeff Bezos " "Bernard Arnault & family " "Bill Gates " ...
##  $ Networth : num  219 171 158 129 118 111 107 106 91.4 90.7 ...
##  $ Age      : int  50 58 73 66 91 49 48 77 66 64 ...
##  $ Country  : chr  "United States" "United States" "France" "United States" ...
##  $ Source   : chr  "Tesla, SpaceX" "Amazon" "LVMH" "Microsoft" ...
##  $ Industry : chr  "Automotive " "Technology " "Fashion & Retail " "Technology " ...
##  $ continent: chr  "Americas" "Americas" "Europe" "Americas" ...
```

BUBBLE CHART ANALYSIS

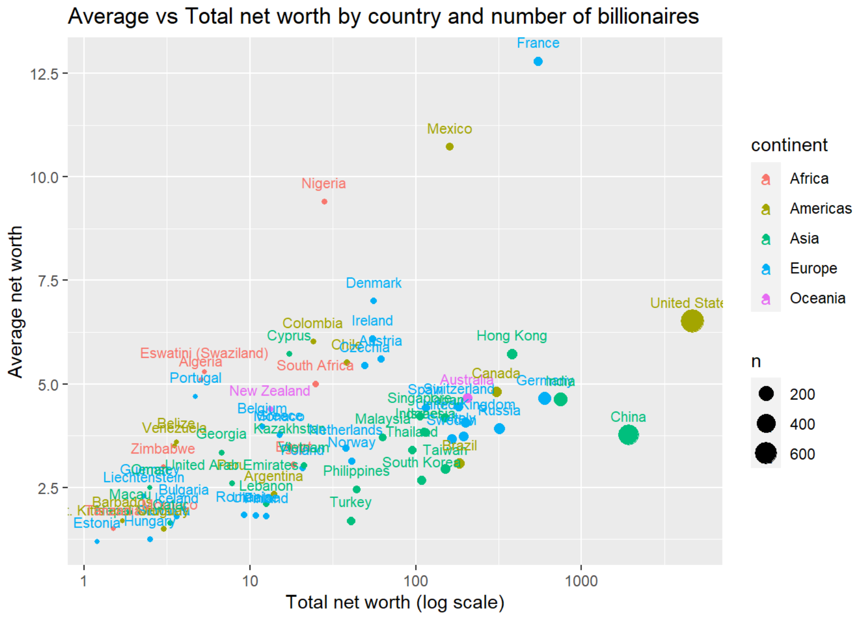
Net Worth X Number X Industry

```
by_indstry <- baires %>%
  group_by(Industry) %>%
  summarize(sum_nw = sum(Networth), avg_nw = mean(Networth), n = n()) %>%
  arrange(sum_nw)

ggplot(by_indstry, aes(sum_nw, n, size = avg_nw, color = Industry)) +
  geom_jitter() +
  labs(title = "Billionaires by industry, net worth, and count",
       x = "Number of billionaires",
       y = "Total net worth (log scale)"
  ) +
  scale_y_log10() +
  geom_text(aes(label=Industry, size = 4), vjust=-1)
```



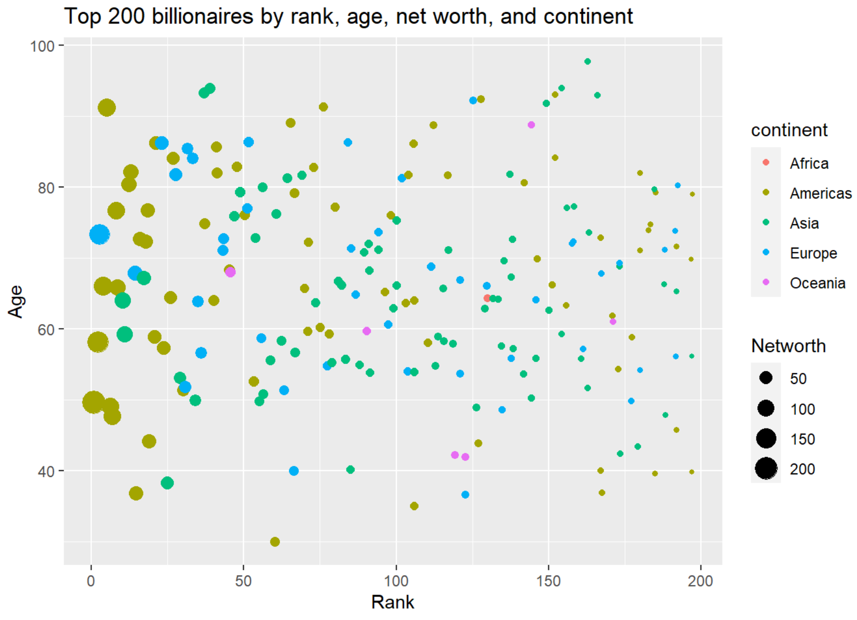
Net Worth X Number X Country X Continent

```
by_cntry <- baires %>%
  group_by(Country) %>%
  summarize(sum_nw = sum(Networth), avg_nw = mean(Networth), n = n(), continent = unique(continent)) %>%
  arrange(sum_nw)

ggplot(by_cntry, aes(sum_nw, avg_nw, size = n, color = continent)) +
  geom_point() +
  labs(title = "Average vs Total net worth by country and number of billionaires",
       x = "Total net worth (log scale)",
       y = "Average net worth"
       ) +
  scale_x_log10() +
  geom_text(aes(label=Country, size = 100, vjust=-1.25))
```



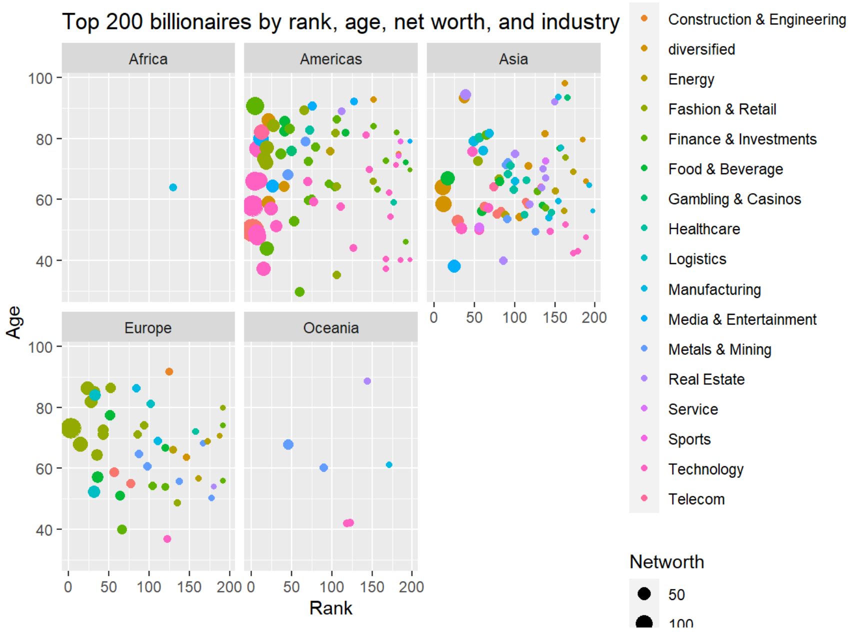
Rank X Age X Net Worth X Industry

```
baires %>%
  filter(Rank <= 200) %>%
    ggplot(aes(Rank, Age, size = Networth, color = continent )) +
    geom_jitter() +
    labs(title = "Top 200 billionaires by rank, age, net worth, and continent",
       x = "Rank",
       )
```



Rank X Age X Net Worth X Industry X Continent

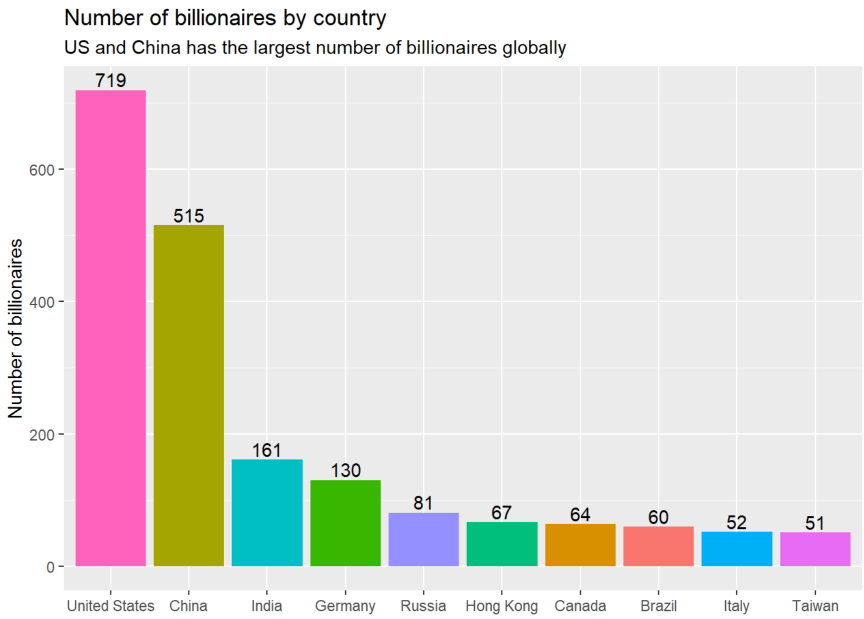
```
baires %>%
  filter(Rank <= 200) %>%
    ggplot(aes(Rank, Age, size = Networth, color = Industry )) +
    geom_jitter() +
    facet_wrap(~continent) +
    labs(title = "Top 200 billionaires by rank, age, net worth, and industry",
       x = "Rank",
       )
```



COUNTRY ANALYSIS

Column: Number of billionaires by country

```
num_by_cntry <- baires %>%
  group_by(Country) %>%
  summarize(n = n()) %>%
  arrange(desc(n))

total_baires <- nrow(baires)

per_by_cntry <- num_by_cntry %>%
  mutate(percentage = num_by_cntry$n/total_baires*100)

top10_per_by_cntry <- head(per_by_cntry, n = 10)

ggplot(top10_per_by_cntry, aes(reorder(Country, -n), n, fill = Country)) +
  geom_col() +
  theme(legend.position = "none",
        axis.title.x=element_blank()) +
  labs(title = "Number of billionaires by country",
       subtitle = "US and China has the largest number of billionaires globally",
       y = "Number of billionaires"
       ) +
  geom_text(aes(label=round(n,1)), vjust=-0.25)
```

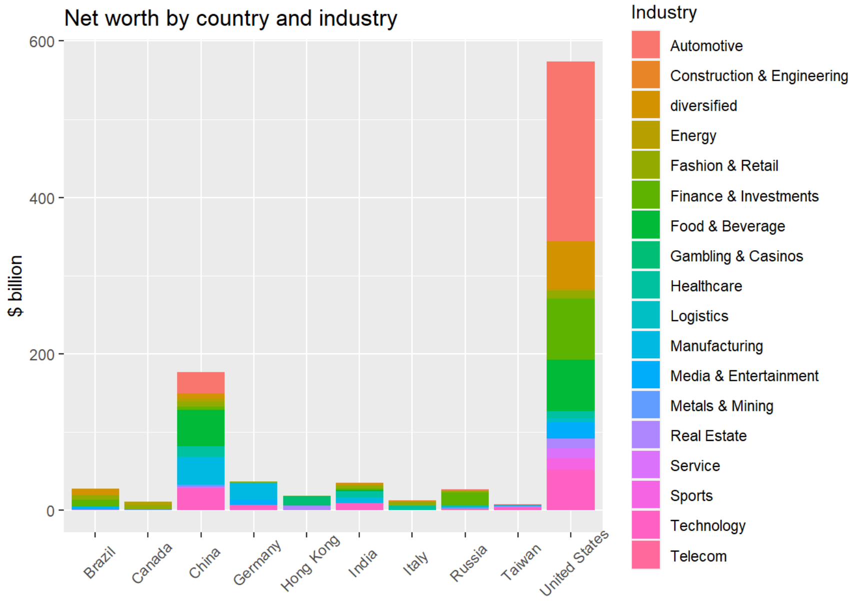


Net Worth by Country x Industry

```
top10c <- top10_per_by_cntry$Country

nw_by_cntry_x_indstry <- baires %>%
  filter(Country == top10c) %>%
  group_by(Country, Industry) %>%
  summarize(sum_nw = sum(Networth))
```

```
## `summarise()` has grouped output by 'Country'. You can override using the
## `.groups` argument.
```

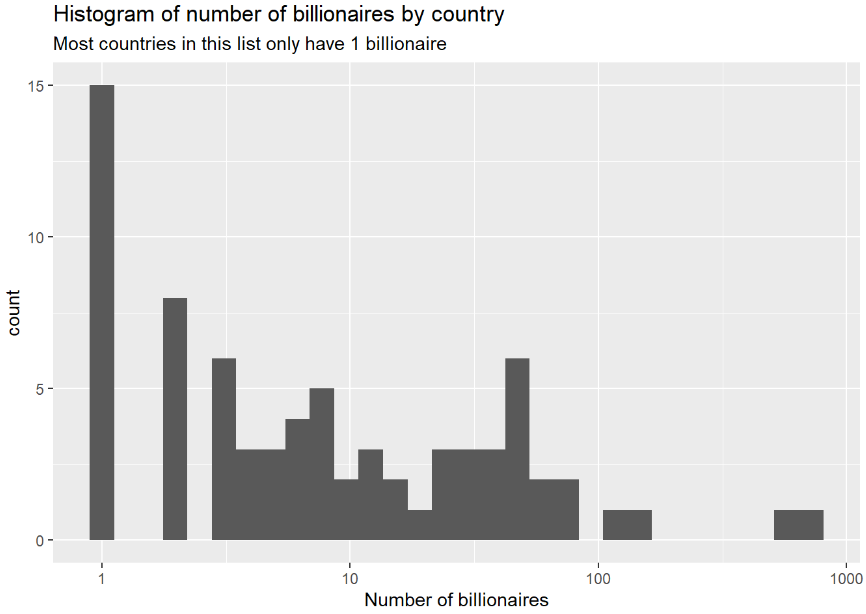
```
ggplot(nw_by_cntry_x_indstry, aes(Country, sum_nw, fill = Industry)) +
  geom_col() +
  theme(axis.title.x=element_blank()) +
  labs(title = "Net worth by country and industry",
       y = "$ billion"
       ) +
  theme(axis.text.x = element_text(angle = 45, vjust=0.75))
```



Histogram: Number of billionaires by country

```
ggplot(num_by_cntry, aes(n)) +
  geom_histogram() +
  theme(legend.position = "none") +
  labs(title = "Histogram of number of billionaires by country",
       subtitle = "Most countries in this list only have 1 billionaire",
       x = "Number of billionaires"
       ) +
  scale_x_log10()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

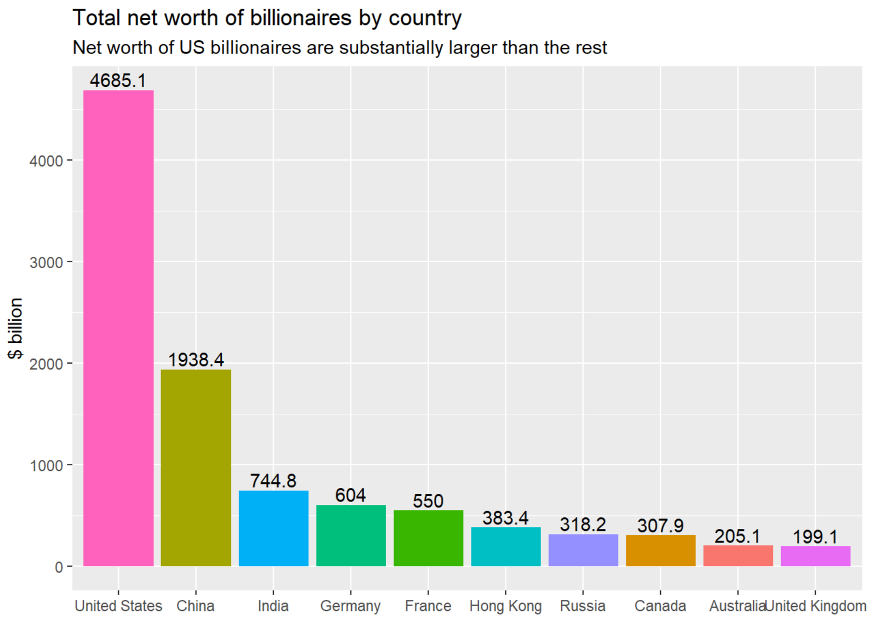


Histogram of number of billionaires by country

Most countries in this list only have 1 billionaire

Column: Net worth by country

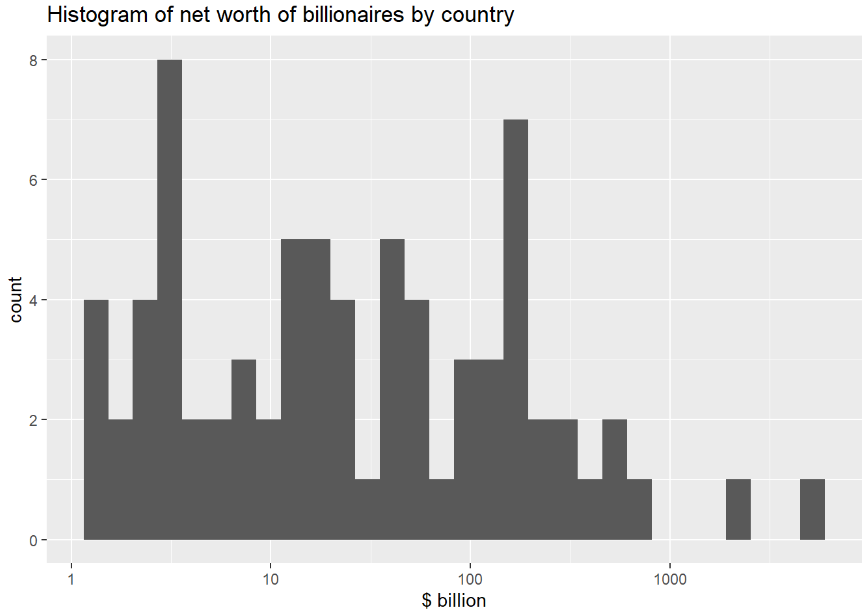
```
nw_by_cntry <- baires %>%
  group_by(Country) %>%
  summarize(sum_nw = sum(Networth)) %>%
  arrange(desc(sum_nw))

total_nw <- sum(baires$Networth)

pernw_by_cntry <- nw_by_cntry %>%
  mutate(percentage = nw_by_cntry$sum_nw/total_nw*100)

top10_pernw_by_cntry <- head(pernw_by_cntry, n = 10)

ggplot(top10_pernw_by_cntry, aes(reorder(Country, -sum_nw), sum_nw, fill = Country)) +
  geom_col() +
  theme(legend.position = "none",
        axis.title.x=element_blank()) +
  labs(title = "Total net worth of billionaires by country",
       subtitle = "Net worth of US billionaires are substantially larger than the rest",
       y = "$ billion"
       ) +
  geom_text(aes(label=round(sum_nw,1)), vjust=-0.25)
```



Histogram: Net worth of billionaires by country

```
ggplot(nw_by_cntry, aes(sum_nw)) +
  geom_histogram() +
  theme(legend.position = "none") +
  labs(title = "Histogram of net worth of billionaires by country",
       x = "$ billion"
       ) +
  scale_x_log10()
```

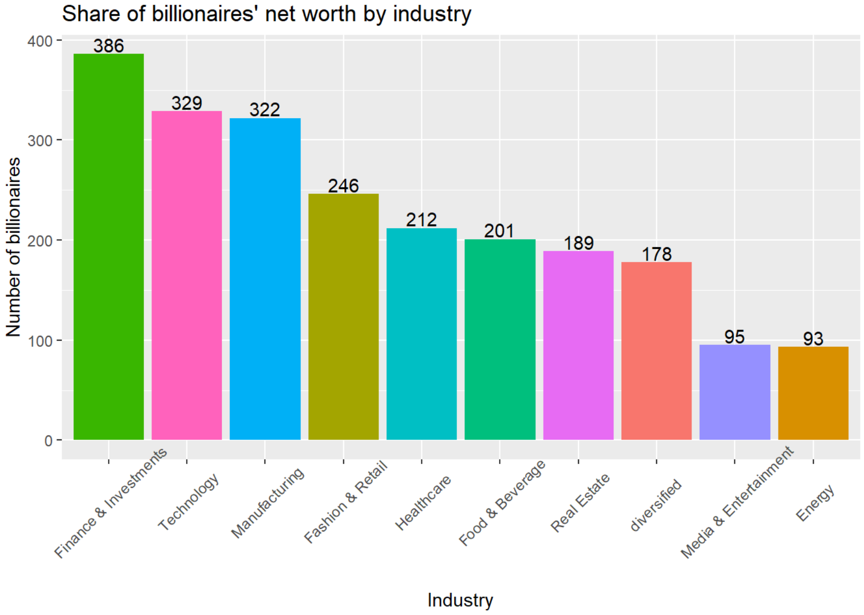
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



INDUSTRY ANALYSIS

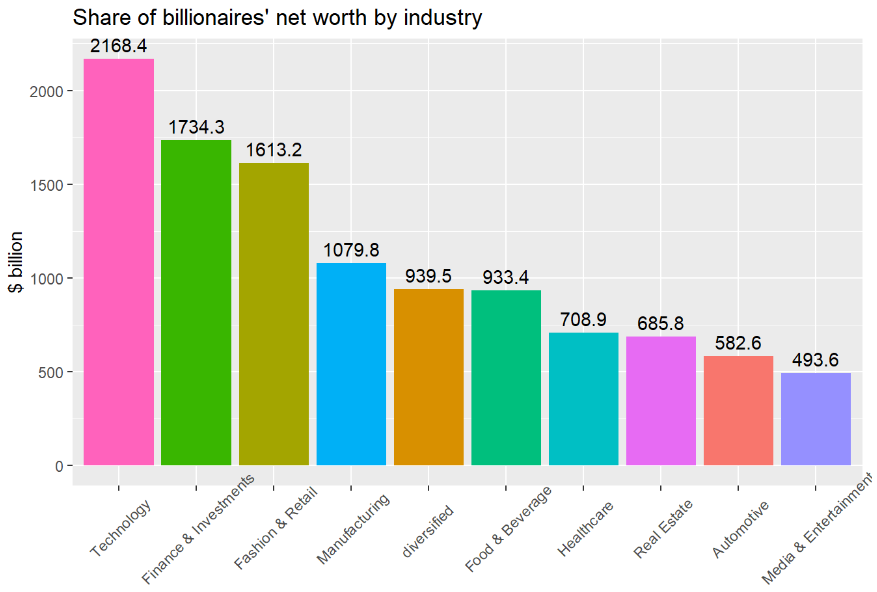
Column: Number of billionaires by industry

```
no_by_indstry <- baires %>%
  group_by(Industry) %>%
  summarize(n = n()) %>%
  arrange(desc(n))

perno_by_indstry <- no_by_indstry %>%
  mutate(percentage = no_by_indstry$n/total_baires*100)

top10_perno_by_indstry <- head(perno_by_indstry, n = 10)

ggplot(top10_perno_by_indstry, aes(reorder(Industry, -n), n, fill = Industry)) +
  geom_col() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, vjust=0.75)) +
  labs(title = "Share of billionaires' net worth by industry",
       x = "Industry",
       y = "Number of billionaires"
       ) +
  geom_text(aes(label=round(n,1)), vjust=-0.15)
```



Column: Net worth by industry

```
nw_by_indstry <- baires %>%
  group_by(Industry) %>%
  summarize(sum_nw = sum(Networth)) %>%
  arrange(desc(sum_nw))

total_nw <- sum(baires$Networth)

pernw_by_indstry <- nw_by_indstry %>%
  mutate(percentage = nw_by_indstry$sum_nw/total_nw*100)

top10_pernw_by_indstry <- head(pernw_by_indstry, n = 10)

ggplot(top10_pernw_by_indstry, aes(reorder(Industry, -sum_nw), sum_nw, fill = Industry)) +
  geom_col() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, vjust=0.75),
        axis.title.x=element_blank()) +
  labs(title = "Share of billionaires' net worth by industry",
       y = "$ billion"
       ) +
  geom_text(aes(label=round(sum_nw,1)), vjust=-0.5)
```

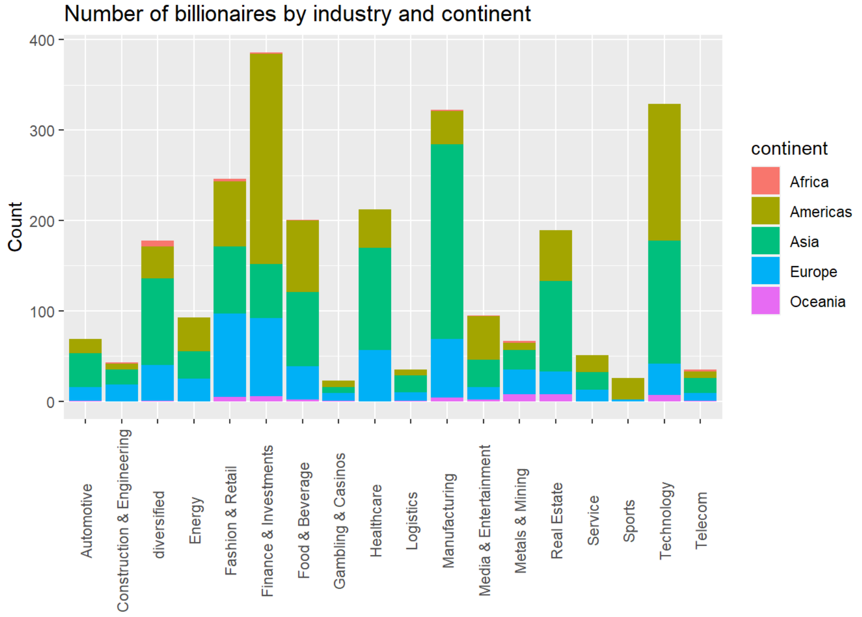


Column: Number of billionaires by industry and continent

```
num_by_indstry_x_continent <- baires %>%
  group_by(Industry, continent) %>%
  summarize(n = n())
```

```
## `summarise()` has grouped output by 'Industry'. You can override using the
## `.groups` argument.
```

```
ggplot(num_by_indstry_x_continent, aes(Industry, n, fill = continent)) +
  geom_col() +
  theme(axis.title.x=element_blank()) +
  labs(title = "Number of billionaires by industry and continent",
       y = "Count"
       ) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5))
```



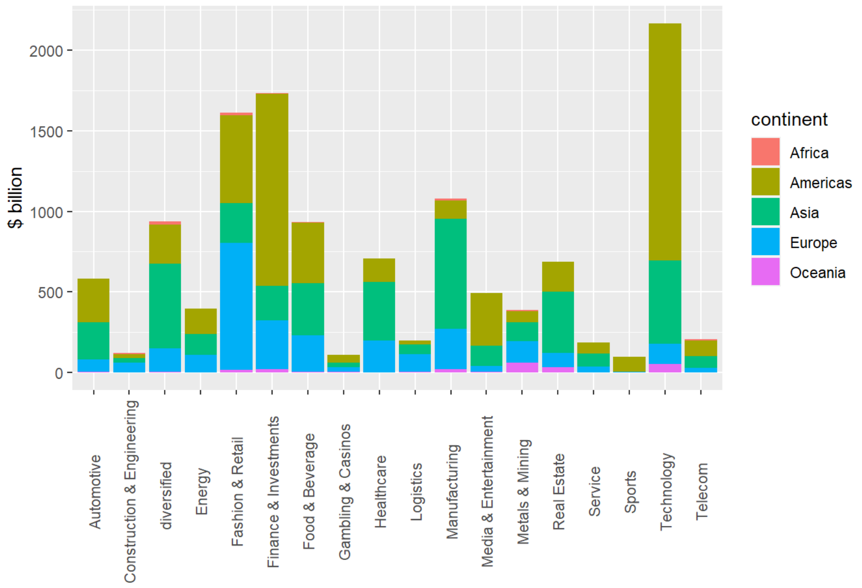
Column: Net worth by industry and continent

```
nw_by_indstry_x_continent <- baires %>%
  group_by(Industry, continent) %>%
  summarize(sum_nw = sum(Networth))
```

```
## `summarise()` has grouped output by 'Industry'. You can override using the
## `.groups` argument.
```

```
ggplot(nw_by_indstry_x_continent, aes(Industry, sum_nw, fill = continent)) +
  geom_col() +
  theme(axis.title.x=element_blank()) +
  labs(title = "Net worth of billionaires by industry and continent",
       y = "$ billion"
       ) +
  theme(axis.text.x = element_text(angle = 90, vjust=0.5))
```

Net worth of billionaires by industry and continent



AGE ANALYSIS

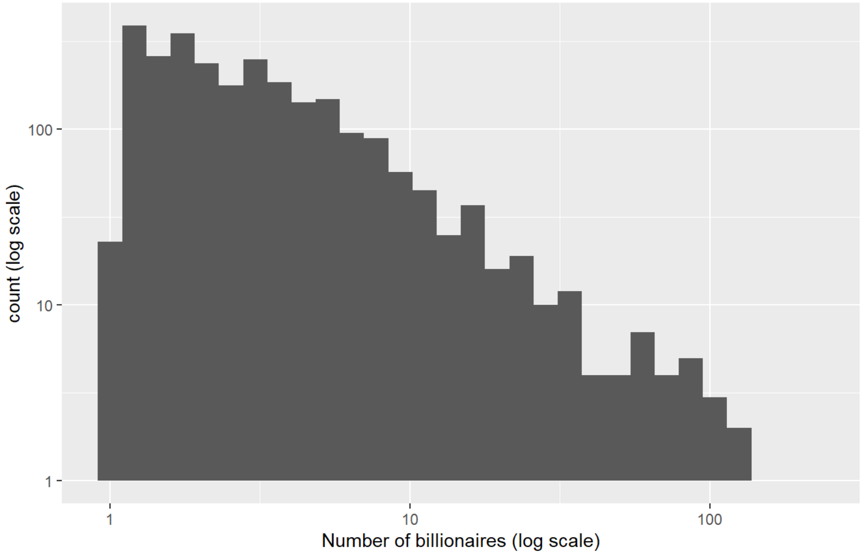
Histogram

```
ggplot(baires, aes(Networth)) +
  geom_histogram() +
  theme(legend.position = "none") +
  labs(title = "Histogram of net worth of billionaires by country",
       subtitle = "Most billionaies have around $1 - $2 billion",
       x = "Number of billionaires (log scale)",
       y = "count (log scale)"
       ) +
  scale_x_log10() +
  scale_y_log10()
```

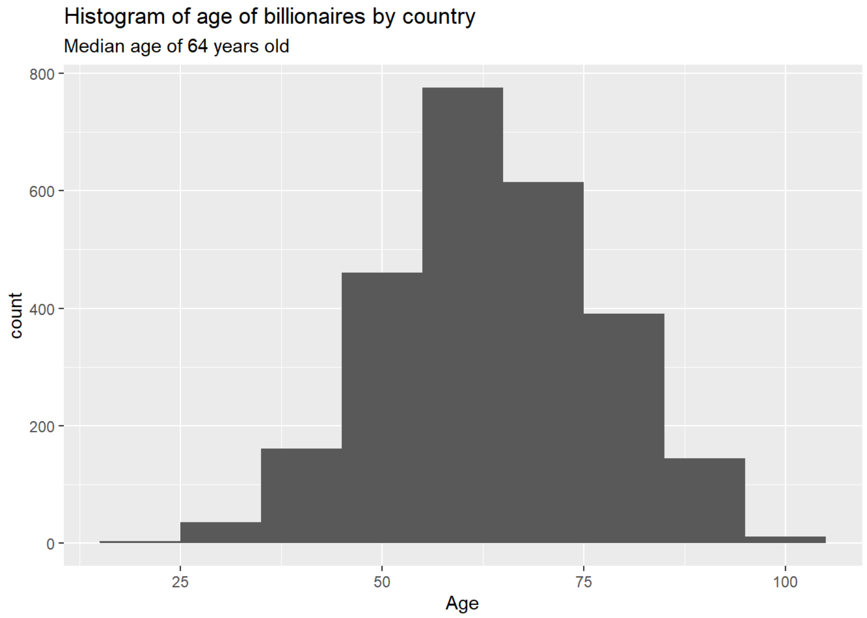
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Histogram of net worth of billionaires by country

Most billionaies have around $1 - $2 billion



```
ggplot(baires, aes(Age)) +
  geom_histogram(binwidth = 10) +
  labs(title = "Histogram of age of billionaires by country",
       subtitle = "Median age of 64 years old",
       x = "Age"
       )
```

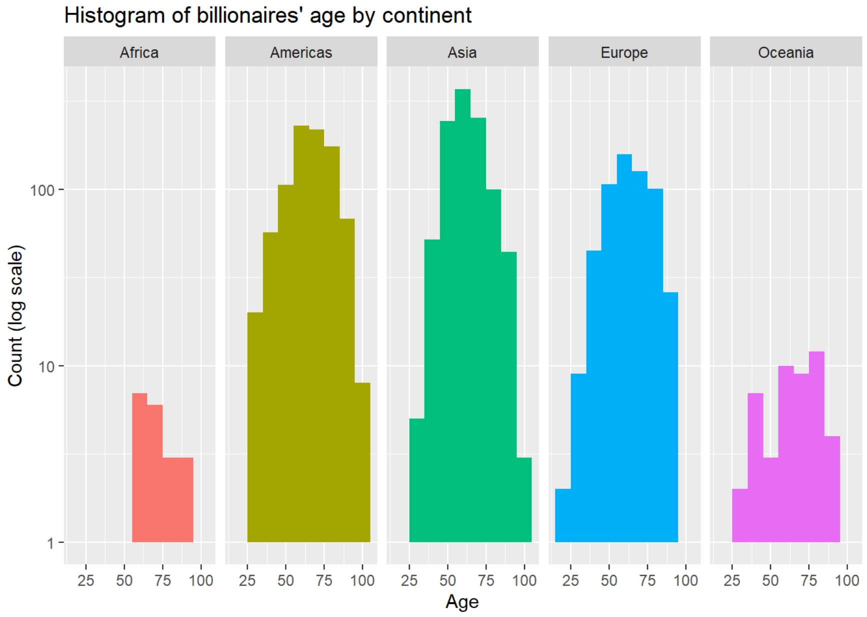


Age by Continent

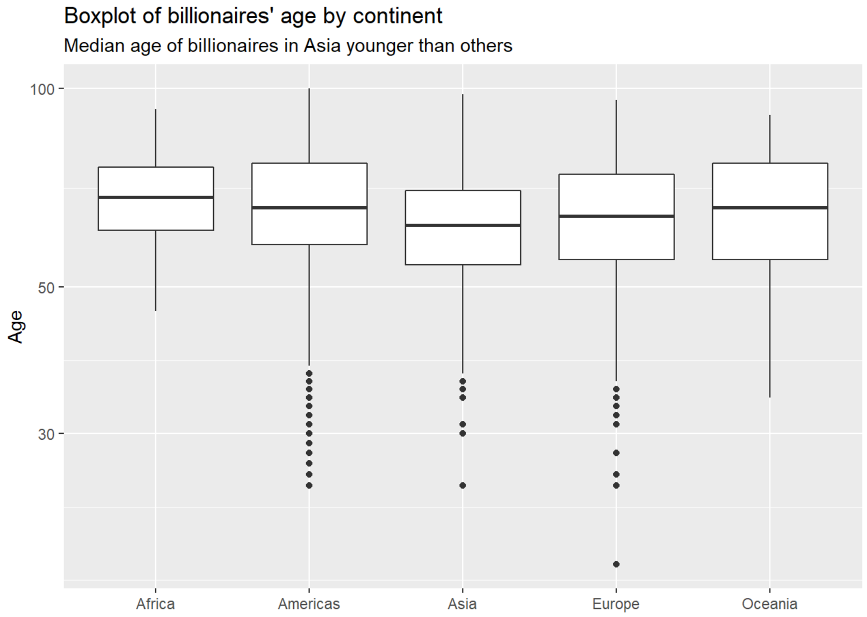
```
ggplot(baires, aes(Age, fill = continent)) +
  geom_histogram(binwidth = 10) +
  facet_grid(~continent) +
  theme(legend.position = "none") +
  scale_y_log10() +
  labs(title = "Histogram of billionaires' age by continent",
       x = "Age",
       y = "Count (log scale)"
       )
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



```
ggplot(baires, aes(continent, Age)) +
  geom_boxplot() +
  theme(legend.position = "none",
        axis.title.x = element_blank()) +
  scale_y_log10() +
  labs(title = "Boxplot of billionaires' age by continent",
       subtitle = "Median age of billionaires in Asia younger than others",
       y = "Age"
       )
```

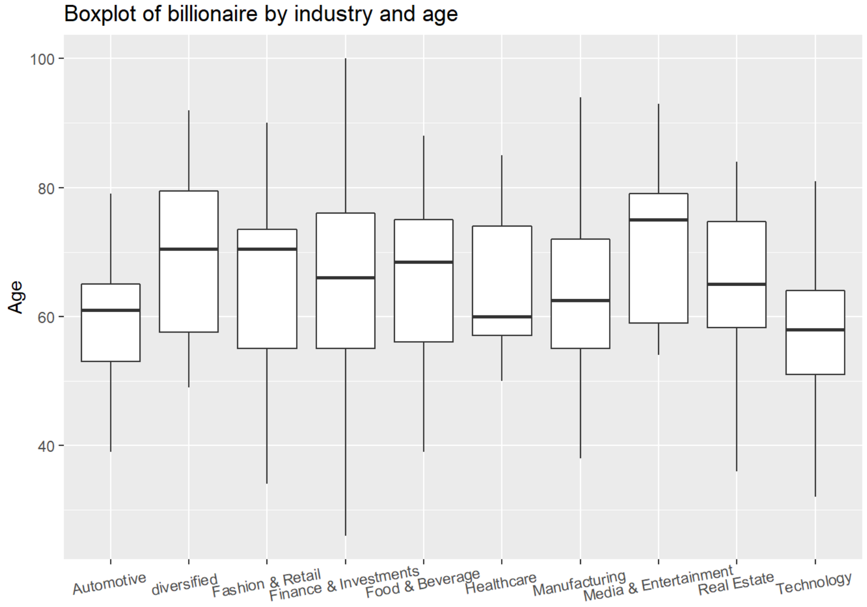


Boxplot of billionaires' age by continent

Median age of billionaires in Asia younger than others

##Age by Industry

```
top10i <- top10_pernw_by_indstry$Industry

baires %>%
  filter(Industry == top10i) %>%
  ggplot(aes(Industry, Age)) +
  geom_boxplot() +
  labs(title = "Boxplot of billionaire by industry and age",
       x = "Industry",
       y = "Age"
       ) +
  theme(axis.text.x = element_text(angle = 10, vjust=0.75),
        axis.title.x = element_blank())
```



Boxplot of billionaire by industry and age

REGRESSION

Regression

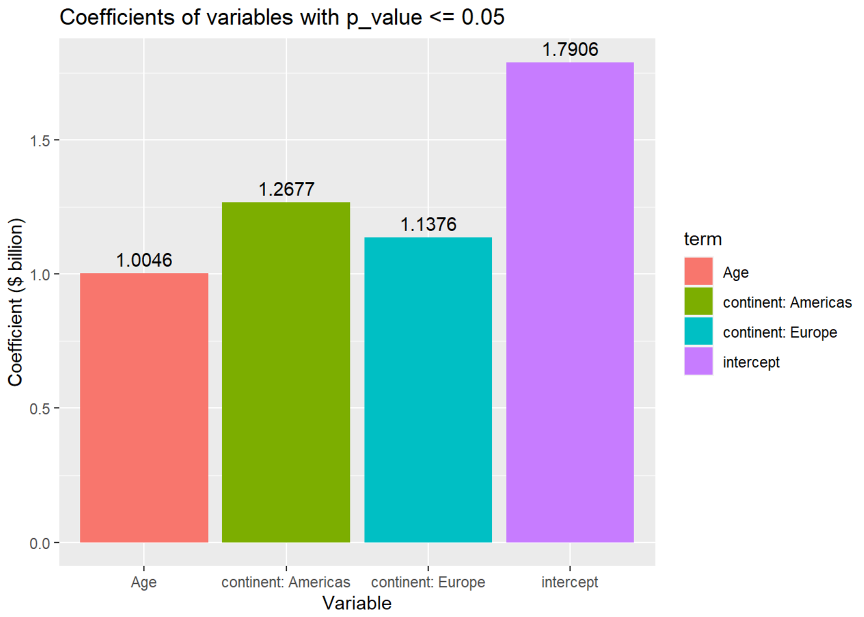
```
baires$continent <- factor(baires$continent, ordered = FALSE)

baires$continent <- relevel(baires$continent, ref = "Asia")

model <- lm(log10(Networth) ~ Age + continent, data = baires)
get_regression_table(model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
intercept	0.253	0.035	7.287	0.000	0.185	0.321
Age	0.002	0.001	4.696	0.000	0.001	0.004
continent: Africa	0.110	0.080	1.387	0.166	-0.046	0.266
continent: Americas	0.103	0.016	6.403	0.000	0.072	0.135
continent: Europe	0.056	0.018	3.050	0.002	0.020	0.091
continent: Oceania	0.075	0.053	1.422	0.155	-0.028	0.178

6 rows

Regression - Coefficients

```
regtab <- get_regression_table(model)

regtab %>%
  filter(p_value <=0.1) %>%
  ggplot(aes(term, 10^(estimate), fill = term)) +
    geom_col() +
    labs(title = "Coefficients of variables with p_value <= 0.05",
       x = "Variable",
       y = "Coefficient ($ billion)"
       ) +
    geom_text(aes(label=round(10^estimate,4), vjust=-.5))
```



# QUICK FACTS

NETWORK RANGE OF BILLIONAIRES - MAX, MIN, MEAN

```
nw_descriptive_stats <- c(nrow(baires), max(baires$Networth), min(baires$Networth), mean(baires$Networth), median(baires$Net
worth))
names(nw_descriptive_stats) <- c("Number of billionaires", "Highest net worth", "Lowest net worth", "Average net worth", "Me
dian net worth")
nw_descriptive_stats
```

```
## Number of billionaires      Highest net worth       Lowest net worth
##             2600.00000              219.00000                1.00000
##      Average net worth       Median net worth
##                4.86075                2.40000
```

RICHEST, YOUNGEST, OLDEST, MALAYSIAN BILLIONAIRES

```
richest_baires <- baires %>%
    arrange(desc(Networth)) %>%
    head(n = 10L)

youngest_baires <- baires %>%
    arrange(Age) %>%
    head(n = 10L)

oldest_baires <- baires %>%
    arrange(desc(Age)) %>%
    head(n = 10L)

MY_baires <- baires %>%
    filter(Country == "Malaysia") %>%
    arrange(desc(Networth))


richest_baires
```

	Rank	Name	Networth	A…	Country	Source
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>
1	1	Elon Musk	219.0	50	United States	Tesla, SpaceX
2	2	Jeff Bezos	171.0	58	United States	Amazon
3	3	Bernard Arnault & family	158.0	73	France	LVMH
4	4	Bill Gates	129.0	66	United States	Microsoft
5	5	Warren Buffett	118.0	91	United States	Berkshire Hathaway
6	6	Larry Page	111.0	49	United States	Google
7	7	Sergey Brin	107.0	48	United States	Google
8	8	Larry Ellison	106.0	77	United States	software
9	9	Steve Ballmer	91.4	66	United States	Microsoft
10	10	Mukesh Ambani	90.7	64	India	diversified

1-10 of 10 rows | 1-7 of 9 columns

```
youngest_baires
```

	Rank	Name	Networth	A…	Country	Source
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>
1	1292	Kevin David Lehmann	2.4	19	Germany	drugstores
2	1929	Pedro Franceschi	1.5	25	Brazil	fintech
3	1929	Wang Zelong	1.5	25	China	chemicals
4	2190	Alexandra Andresen	1.3	25	Norway	investments
5	1929	Henrique Dubugras	1.5	26	Brazil	fintech
6	2190	Katharina Andresen	1.3	26	Norway	investments
7	1513	Ryan Breslow	2.0	27	United States	e-commerce software
8	1818	Austin Russell	1.6	27	United States	sensors★
9	431	Gary Wang	5.9	28	United States	cryptocurrency exchange
10	637	Gustav Magnar Witzoe	4.5	28	Norway	fish farming

1-10 of 10 rows | 1-7 of 9 columns

```
oldest_baires
```

	R…	Name	Networth	…	Country	Source
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>
1	1645	George Joseph	1.8	100	United States	insurance
2	163	Robert Kuok	11.7	98	Malaysia	palm oil, shipping, property
3	1238	Charles Munger	2.5	98	United States	Berkshire Hathaway
4	1341	David Murdock	2.3	98	United States	Dole, real estate
5	622	Masatoshi Ito	4.6	97	Japan	retail
6	1513	S. Daniel Abraham	2.0	97	United States	Slim-Fast
7	1929	Ana Maria Brescia Cafferata	1.5	97	Peru	mining, banking
8	637	Ted Lerner & family	4.5	96	United States	real estate
9	1645	Stephen Jarislowsky	1.8	96	Canada	money management

	R…	Name	Networth	…	Country	Source	▸
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>	
10	1929	John Farber	1.5	96	United States	chemicals	

1-10 of 10 rows | 1-7 of 9 columns

```
MY_baires
```

Rank	Name	Networth	…	Country	Source	▸
<int>	<chr>	<dbl>	<int>	<chr>	<chr>	
163	Robert Kuok	11.7	98	Malaysia	palm oil, shipping, property	
185	Quek Leng Chan	10.6	80	Malaysia	banking, property	
431	Teh Hong Piow	5.9	92	Malaysia	banking	
460	Ananda Krishnan	5.7	84	Malaysia	telecoms, media, oil-services	
523	Koon Poh Keong	5.2	60	Malaysia	aluminum	
586	Yeow Chor & Yeow Seng Lee	4.8	64	Malaysia	palm oil, property	
1196	Chen Lip Keong	2.6	74	Malaysia	casinos, property, energy	
1445	Lau Cho Kun	2.1	86	Malaysia	palm oil, property	
1513	Kuan Kam Hon & family	2.0	74	Malaysia	rubber gloves	
1513	Lim Kok Thay	2.0	70	Malaysia	casinos	

1-10 of 17 rows | 1-6 of 8 columns Previous **1** 2 Next

AGE RANGE OF BILLIONAIRES - OLDEST, YOUNGEST, AVERAGE AGE

```
age_descriptive_stats <- c(max(baires$Age), min(baires$Age), mean(baires$Age), median(baires$Age))
names(age_descriptive_stats) <- c("Oldest", "Youngest", "Average age", "Median age")
age_descriptive_stats
```

```
##      Oldest    Youngest Average age  Median age
##   100.00000    19.00000    64.27192    64.00000
```

THE TEN OLDEST AND YOUNGEST BILLIONAIRES IN THE WORLD

```
oldest_baires <- baires %>%
    arrange(desc(Age)) %>%
    head(n = 10L)

youngest_baires <- baires %>%
    arrange(Age) %>%
    head(n = 10L)

oldest_baires
```

	R…	Name	Networth	…	Country	Source	▸
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>	
1	1645	George Joseph	1.8	100	United States	insurance	
2	163	Robert Kuok	11.7	98	Malaysia	palm oil, shipping, property	
3	1238	Charles Munger	2.5	98	United States	Berkshire Hathaway	
4	1341	David Murdock	2.3	98	United States	Dole, real estate	
5	622	Masatoshi Ito	4.6	97	Japan	retail	
6	1513	S. Daniel Abraham	2.0	97	United States	Slim-Fast	
7	1929	Ana Maria Brescia Cafferata	1.5	97	Peru	mining, banking	
8	637	Ted Lerner & family	4.5	96	United States	real estate	
9	1645	Stephen Jarislowsky	1.8	96	Canada	money management	
10	1929	John Farber	1.5	96	United States	chemicals	

1-10 of 10 rows | 1-7 of 9 columns

```
youngest_baires
```

	Rank	Name	Networth	A…	Country	Source	▸
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>	

	Rank	Name	Networth	A…	Country	Source	▸
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>	
1	1292	Kevin David Lehmann	2.4	19	Germany	drugstores	
2	1929	Pedro Franceschi	1.5	25	Brazil	fintech	
3	1929	Wang Zelong	1.5	25	China	chemicals	
4	2190	Alexandra Andresen	1.3	25	Norway	investments	
5	1929	Henrique Dubugras	1.5	26	Brazil	fintech	
6	2190	Katharina Andresen	1.3	26	Norway	investments	
7	1513	Ryan Breslow	2.0	27	United States	e-commerce software	
8	1818	Austin Russell	1.6	27	United States	sensors★	
9	431	Gary Wang	5.9	28	United States	cryptocurrency exchange	
10	637	Gustav Magnar Witzoe	4.5	28	Norway	fish farming	

1-10 of 10 rows | 1-7 of 9 columns

SOURCE OF WEALTH

```
baires %>%
  group_by(Source) %>%
  summarize(sum_nw = sum(Networth)) %>%
  arrange(desc(sum_nw))
```

Source	sum_nw
<chr>	<dbl>
real estate	573.80
diversified	382.00
investments	358.30
software	289.70
pharmaceuticals	284.40
hedge funds	271.60
Google	260.90
Walmart	238.00
Microsoft	232.40
Tesla, SpaceX	219.00

1-10 of 895 rows   Previous **1** 2 3 4 5 6 ... 90 Next