

# Data\_Analysis

Servando De La Garza

3/26/2022

Taking a look at the dataset

```
billionaire.data = read.csv("C:\\Users\\Servando\\Downloads\\archive (34)\\Billionaire.csv")
head(billionaire.data)
```

```
##           Name NetWorth      Country      Source Rank Age
## 1      Jeff Bezos  $177 B United States      Amazon   1  57
## 2      Elon Musk  $151 B United States  Tesla, SpaceX   2  49
## 3 Bernard Arnault & family $150 B      France      LVMH    3  72
## 4      Bill Gates  $124 B United States      Microsoft   4  65
## 5    Mark Zuckerberg   $97 B United States      Facebook   5  36
## 6    Warren Buffett   $96 B United States Berkshire Hathaway 6  90
##           Industry
## 1      Technology
## 2      Automotive
## 3    Fashion & Retail
## 4      Technology
## 5      Technology
## 6 Finance & Investments
```

Checking for type errors and fixing them with casting.(Data Cleaning)

```
require("readr")
```

```
## Loading required package: readr
```

```
#Checking types
```

```
typeof(billionaire.data$NetWorth)
```

```
## [1] "character"
```

```
typeof(billionaire.data$Age)
```

```
## [1] "character"
```

```
#Changing types to appropriate ones
```

```
billionaire.data$NetWorth <- as.numeric(parse_number(billionaire.data$NetWorth))
```

```
billionaire.data$Age <- as.numeric(billionaire.data$Age)
```

```
## Warning: NAs introduced by coercion
```

```
#Changing column name
colnames(billionaire.data)[2]<-"Net_Worth_Billions"
#Checking dataframe
head(billionaire.data)
```

```
##           Name Net_Worth_Billions      Country      Source
## 1      Jeff Bezos           177 United States      Amazon
## 2      Elon Musk           151 United States  Tesla, SpaceX
## 3 Bernard Arnault & family    150      France      LVMH
## 4      Bill Gates           124 United States      Microsoft
## 5      Mark Zuckerberg          97 United States      Facebook
## 6      Warren Buffett          96 United States Berkshire Hathaway
##   Rank Age      Industry
## 1    1  57      Technology
## 2    2  49      Automotive
## 3    3  72  Fashion & Retail
## 4    4  65      Technology
## 5    5  36      Technology
## 6    6  90 Finance & Investments
```

How does the five number summary look like for net worth? What's the standard deviation?

```
summary(billionaire.data$Net_Worth_Billions)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.000   1.500   2.300   4.749   4.200 177.000
```

```
sd(na.omit(billionaire.data$Net_Worth_Billions))
```

```
## [1] 9.615358
```

Let's figure out how the distribution looks like.

```
require("ggplot2")
```

```
## Loading required package: ggplot2
```

```
require("moments")
```

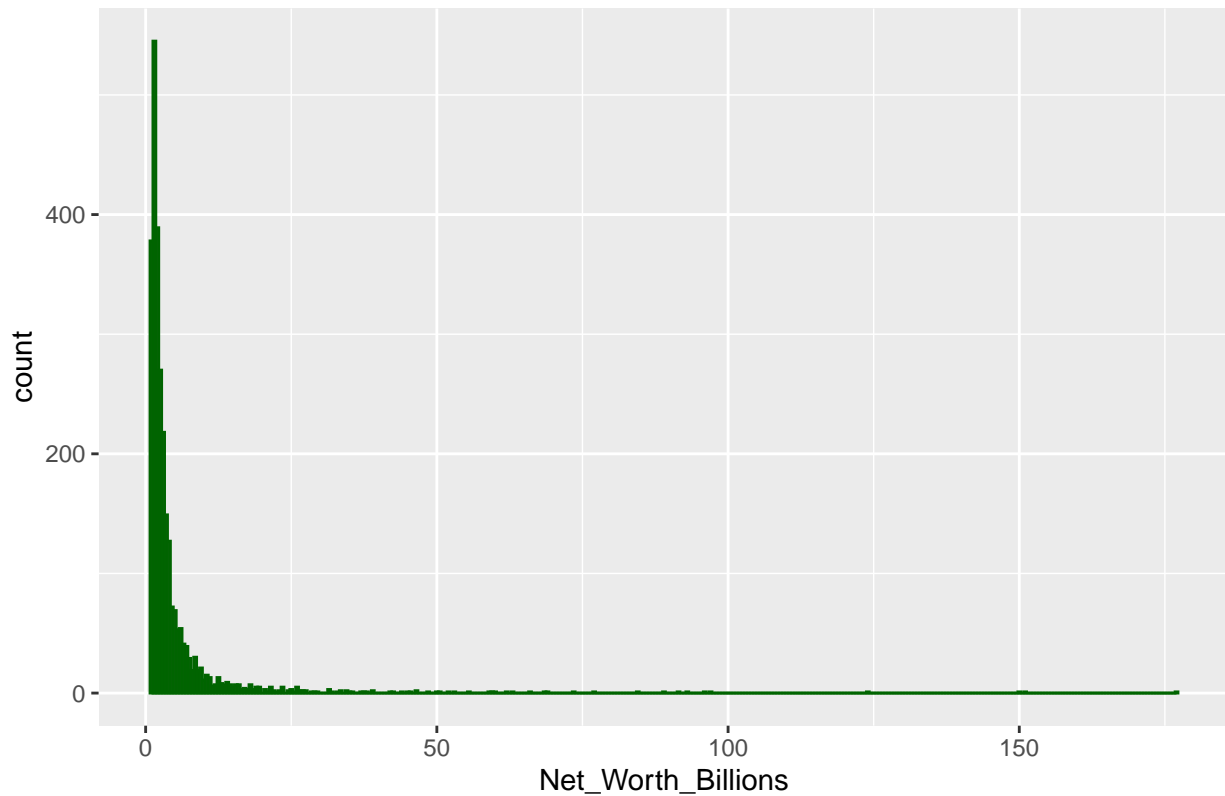
```
## Loading required package: moments
```

```
Networth_billions <- billionaire.data$Net_Worth_Billions
skewness(billionaire.data$Net_Worth_Billions)
```

```
## [1] 8.671725
```

```
ggplot ( data = billionaire.data)+geom_histogram(mapping = aes(x =Net_Worth_Billions ),binwidth = 0.5,color = "red",fill = "red")
```

Net Worth Frequency Distribution



What's the number of billionaires per industry?

```
require("dplyr")
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
```

```
## 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
```

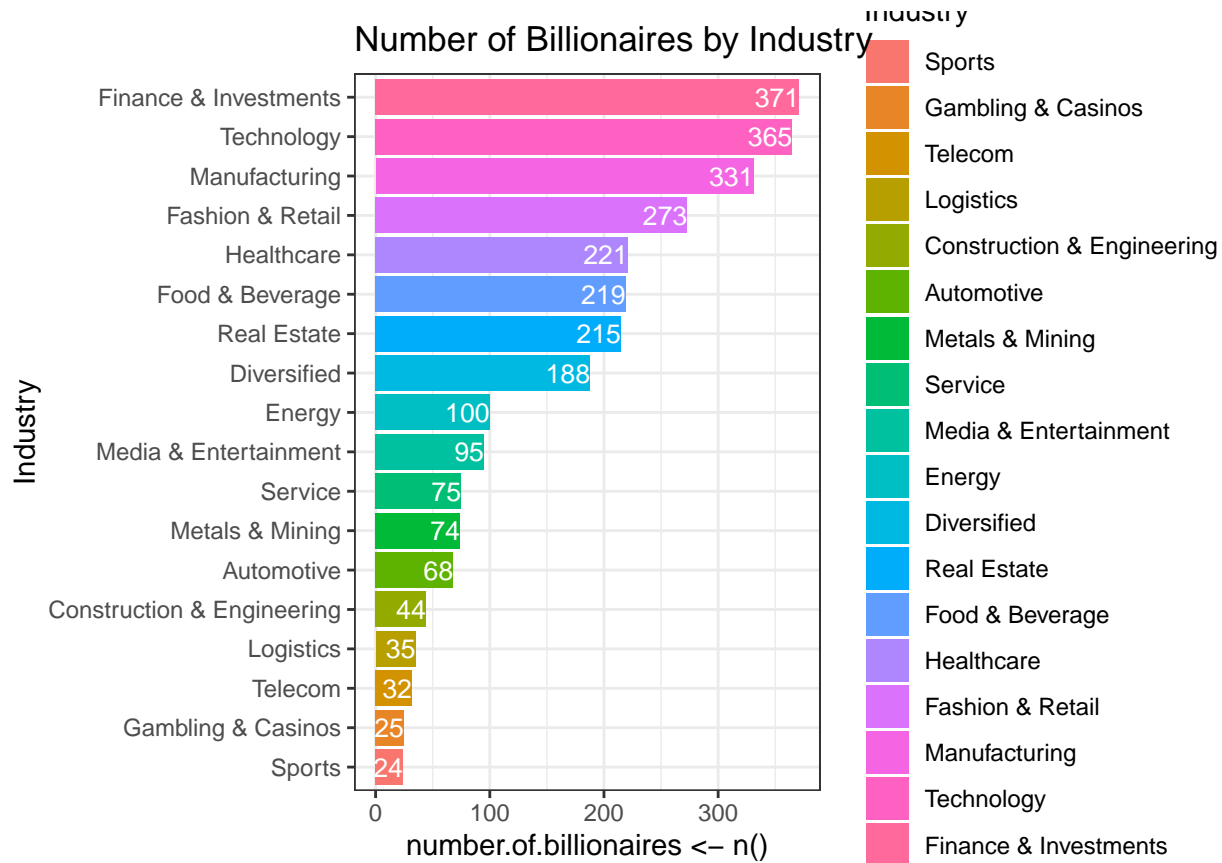
```
require("wesanderson")
```

```
## Loading required package: wesanderson
```

```

by.industry <- billionaire.data %>%
group_by(Industry) %>%
summarise(number.of.billionaires <- n()
)
by.industry$Industry<- factor(by.industry$Industry,levels = by.industry$Industry[order(by.industry$`number.of.billionaires <- n()`)])
par(mar=c(19,4,4,1))
industry.plot <- ggplot(data = by.industry, aes(x = Industry, y = `number.of.billionaires <- n()`, fill = by.industry$Industry))
  geom_text(aes(label = `number.of.billionaires <- n()`, hjust = 1 ,color = "white", size = 3.5))+
  theme_minimal()+coord_flip()
industry.plot + theme_bw()

```



Now, let's break it down by country.

```

by.country <- billionaire.data %>%
group_by(Country) %>%
summarise(number.of.billionaires <- n()
)
by.country <- by.country[order(-by.country$`number.of.billionaires <- n()`) ,]
colnames(by.country)[2]<-"Number of Billionaires"
top<- top_n(by.country,5)

```

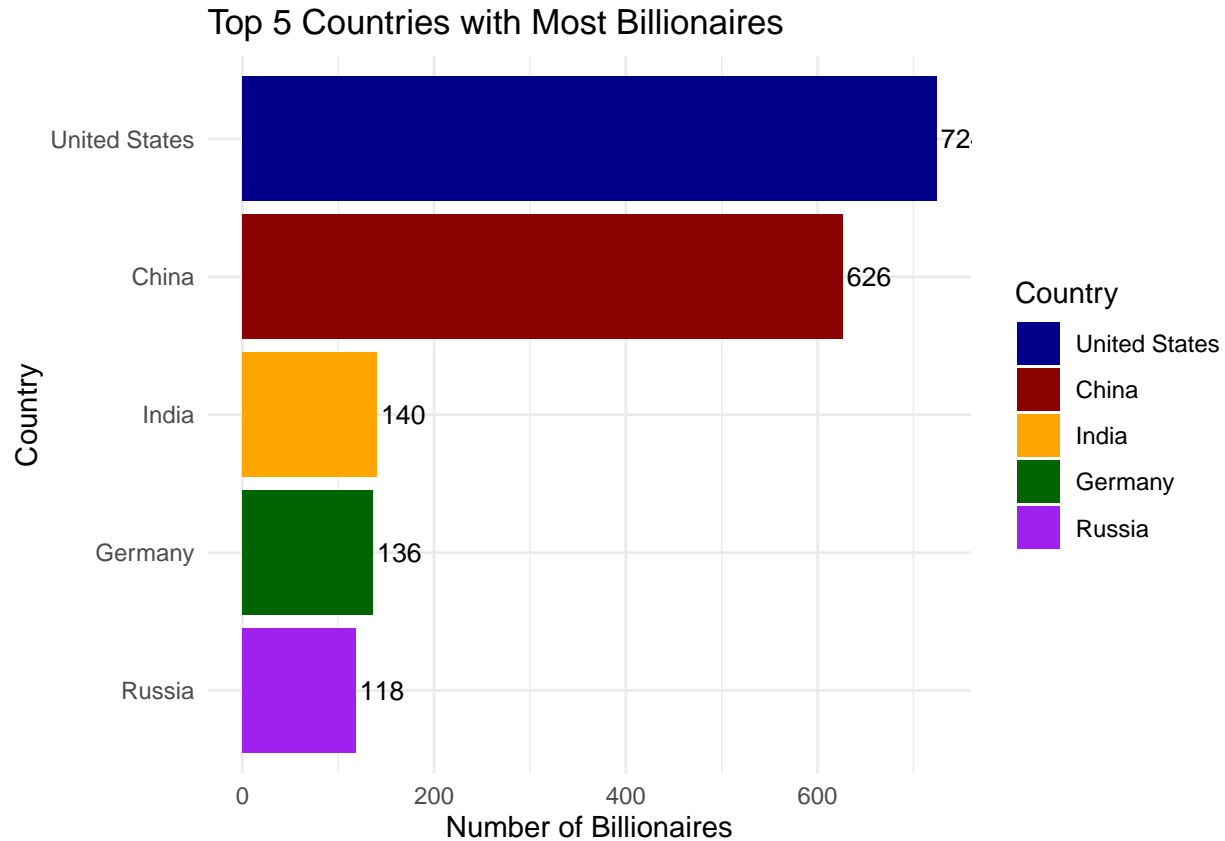
## Selecting by Number of Billionaires

```

top$Country<- factor(top$Country,levels = top$Country[order(top$`Number of Billionaires`)])
group.colors <-c(`United States` = "darkblue", China="Darkred",India = "Orange",Germany="Darkgreen",Russia="Darkred")

```

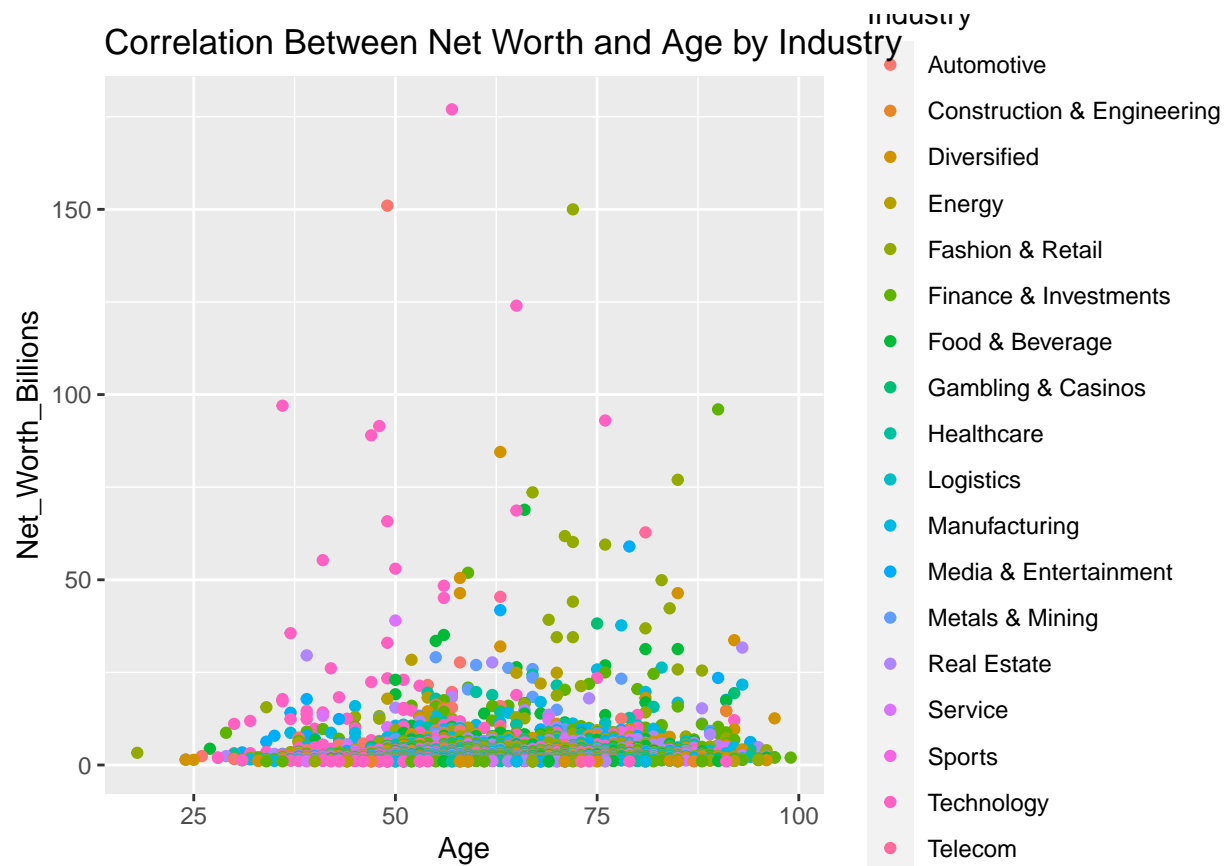
```
ggplot(data = top, aes(x = Country, y = `Number of Billionaires`, fill = Country ))+
  geom_bar(stat = "identity")+ coord_flip() +ggtitle("Top 5 Countries with Most Billionaires") +
  geom_text(aes(label = `Number of Billionaires`), hjust = -.1 , size = 3.5)+
  theme_minimal() + scale_fill_manual(values=group.colors)
```



Let's check for a correlation between age and NetWorth.

```
require("ggplot2")
par(mar=c(10,6,6,1))
ggplot(billionaire.data) + geom_point(mapping = aes(x = Age , y = Net_Worth_Billions ,color=Industry))+
  ggtitle("Correlation Between Net Worth and Age by Industry")
```

## Warning: Removed 79 rows containing missing values (geom\_point).



Let's look at the five number summary of the billionaires' age.

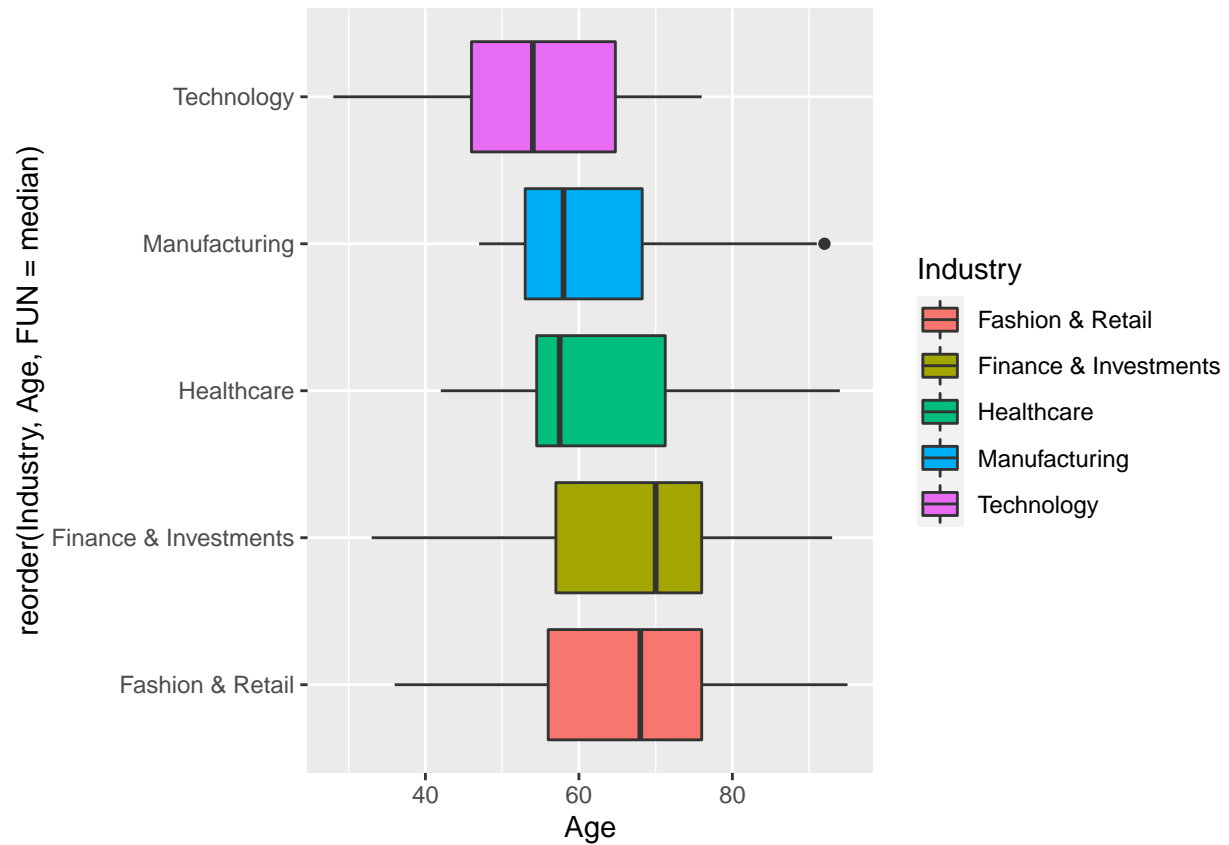
```
summary(billionaire.data$Age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##  18.00   54.00   63.00   63.11   73.00   99.00    79
```

Now, Let's compare the age of billionaires of the 5 industries with most billionaires

```
five.industries <- billionaire.data %>% filter(Industry == c("Finance & Investments", "Technology", "Manufacturing", "Real Estate", "Healthcare"))
agesplot <- ggplot(data = five.industries, mapping = aes(x=reorder(Industry, Age, FUN=median), y = Age, color = Industry)) +
  geom_boxplot() +
  coord_flip()
agesplot
```

```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```



Now, let's see how that comparison it's like for net worth

```
ggplot(data = five.industries , mapping = aes(x=reorder(Industry,Net_Worth_Billions,FUN=median), y = Net_Worth_Billions)) +
  geom_boxplot() +
  coord_flip()
```

