proposal_1

Christina Lyu, Aoi Ogawa, Ahlam

10/2/2018

Group Members:

Christina Lyu, Aoi Ogawa, Ahlam El Mernissi

Title:

The characteristics of billionaires – how to become a billionaire

Purpose:

Most of us wonder who are the billionaires, where are they from and how did they become billionaires. Therefore, we want to look at the characteristics of the world's billionaires, including are they female or male, or are they self made billionaires or did they inherit their wealth. Do such characteristics play a role in determining the net worth of the billionaires worldwide?

Data:

Our data was found on this website. Each row represent a billionare from 1996 to 2016, with columns representing different factors including name, citizenship, net worth of billionaire, whether they are in a developing country(not in the north) or is in the developed country (north) and so on. https://data.opendatasoft.com/explore/dataset/the-billionaire-characteristics-database%40public-us/information/?dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6InRoZS1iaWxsaW9uYWlyZS1jaGFyYWN0ZXJpc3RpY3MtZGF0YW
3D

The dataset presents the source of billionaires wealth and uses it to describe changes in extreme wealth in the United States, Europe, and other advanced countries. The dataset distiguishes whether the wealth is self-made or inherited and idintify the company and industry which it comes from.

Population:

The population of this study is people who are considered as billionaires according to the book Rich People Poor Countries which the data runs from 1996-2014 and the data has been updated to reflect the latest using Forbes. There are 2615 rows which means 2615 people.

Response Variable:

The response variable is the Net.Worth.Billion in billion dollars with unit of 1 billion dollars.

Explanatory Variables:

The explanatory variables for this data set include but are not limited to:

Type of Wealth: the way the billionaire acquired his or her wealth. The possible values could be "inherited", meaning that the billioniare inherited family assets, or "self made finances", meaning that the billionaire started a new company of finance.

Sector: the general industry the billionaire's companies lie in. It is a more generalized grouping of the previous column "Source.of.Wealth" which will not be included in the dataset for analysis because it contains too many different values. Possible values for Sector are "media" or "construction" indicating that the main industry of the billionaire's business is in "media" or "construction".

Region: the area where the billionaire is from. It is the continent of the country given in the previous column "Citizenship". The number of countries in the dataset is too large and including all of them will be overwhelming for the visualization. Possible values for region are "Europe" or "South America".

Gender: the gender of the billionaire. Possible values are "Male" or "Female".

Age: the age of the billionaires in the dataset. It refers to the current age of the billionaire and not the age when he or she first became billionaire. It is a quantitative variable and has values of "41", "83" and others.

Above are some of the variables that could have significant relationship with the amount of dollars a billionaire holds. But it is very possible that there are other variables in the dataset that are significant as well.

Pre-Analysis:

We are interested in how the variables above are related to the net worth of the billionaires. We suspect that there are more males than femeales as billionaires in the dataset, more inherited billionaires than self-made billionaires. We also assume that there are more old people than young and more billionaires in developed countries than developing ones. We would try to justify these assumptions by looking at the distribution of the data points.

Getting the data and processing

```
library(readr)
library(dplyr)
library(ggplot2)

# read the csv file
billionaire <- read.csv("the-billionaire-characteristics-database.csv")

#tidy up the dataset
tidy_billionaire <- subset(billionaire, select = -c(3, 4, 5, 10, 15, 17, 21, 23, 24, 25))

# change the blank cell into a NA
tidy_billionaire[tidy_billionaire == ""] <- NA

without_na <- na.omit(tidy_billionaire)

# excluded NA for gender
na_gender <- subset(tidy_billionaire, !is.na(Gender))

# excluded NA for selfmade and exclude NA for gender
gender_n_Selfmade <- subset(na_gender, !is.na(Selfmade))
```

```r
# create a age group (youth, young_adult, adult, senior)

age_billionare <- gender_n_Selfmade %>%
  mutate(age_group = ifelse((Age >= -42 & Age <= 18 & Age == 0), 'youth', ifelse((Age > 18 & Age <= 35)


# change NA in Political.Connections into 0
gender_n_Selfmade$Political.Connection[is.na(gender_n_Selfmade$Political.Connection)] <- 0


# remove all the na
na_billionare <- na.omit(tidy_billionaire)
# there are only 61 observations when removed all the na from the dataset we removed some columns
```

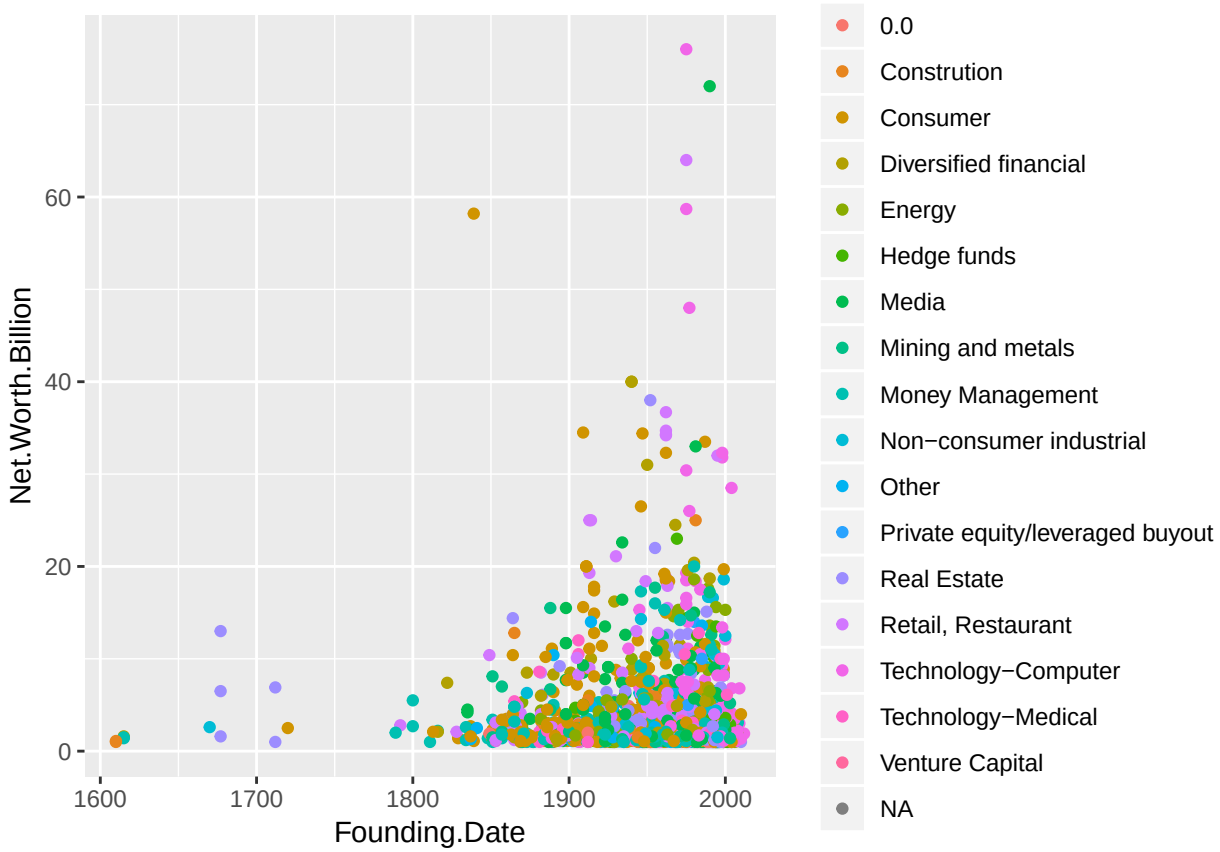
the columns that we deleted

- name
- citizenship
- county code
- Source.of.Wealth
- GDP.Currentus
- Notes
- Notes.2
- Generation.of.Inheretence
- Deflator.1996
- Company

We deleted the columns because they either have a lot of NA values which means only partial of the data could be analyzed or they are repeated fields with other columns in the data set.

Below are some brief examples of our analysis:

```r
# color code by Industry
ggplot(gender_n_Selfmade, aes(x = Founding.Date, y = Net.Worth.Billion, color = Industry)) +
  geom_point()
```



Analysis of graph 1:

For the scatterplot Net.Worth.Billion vs Founding.Date, we examine how net worth (in billions) was affected by different industries over time. For instance, around the 1600s, majority of the billionaires worked in Real Estate. However, as time passed, we begin to see that most billionaires worked in Consumer industry and Technology-Computer industry. This makes sense as we would expect technological advancement to contribute to the wealth of most billionaires. Furthermore, in the late 1900s, we see that billionaires worked in a diverse industry which also makes since due to technological advancement. For instance, technological advancement allowed the Media Industry to expand.

```
# color code by gender and shape by Selfmade or not
ggplot(gender_n_Selfmade, aes(x = Age, y = Net.Worth.Billion, color = Gender, shape = Selfmade)) +
  geom_point()
```

Analysis of graph 2:

For the scatterplot Net.Worth.Billion vs Age, we examine how net worth (in billions) differed among the ages of selfmade male billionaires, selfmade female billionaires, selfmade married couple billionaires, inherited male billionaires, inherited female billionaires, and inherited married couple billionaires. As we expected, most billionaires are male. Furthermore, most billionaires with the highest net worth are selfmade and aged atleast 50.