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.

```
library(readr)
library(dplyr)

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)

na_gender <- subset(tidy_billionaire, !is.na(Gender))

gender_n_Selfmade <- subset(na_gender, !is.na(Selfmade))

sum(is.na(tidy_billionaire$Gender))
```

```
## [1] 3765
```

```
# 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
```

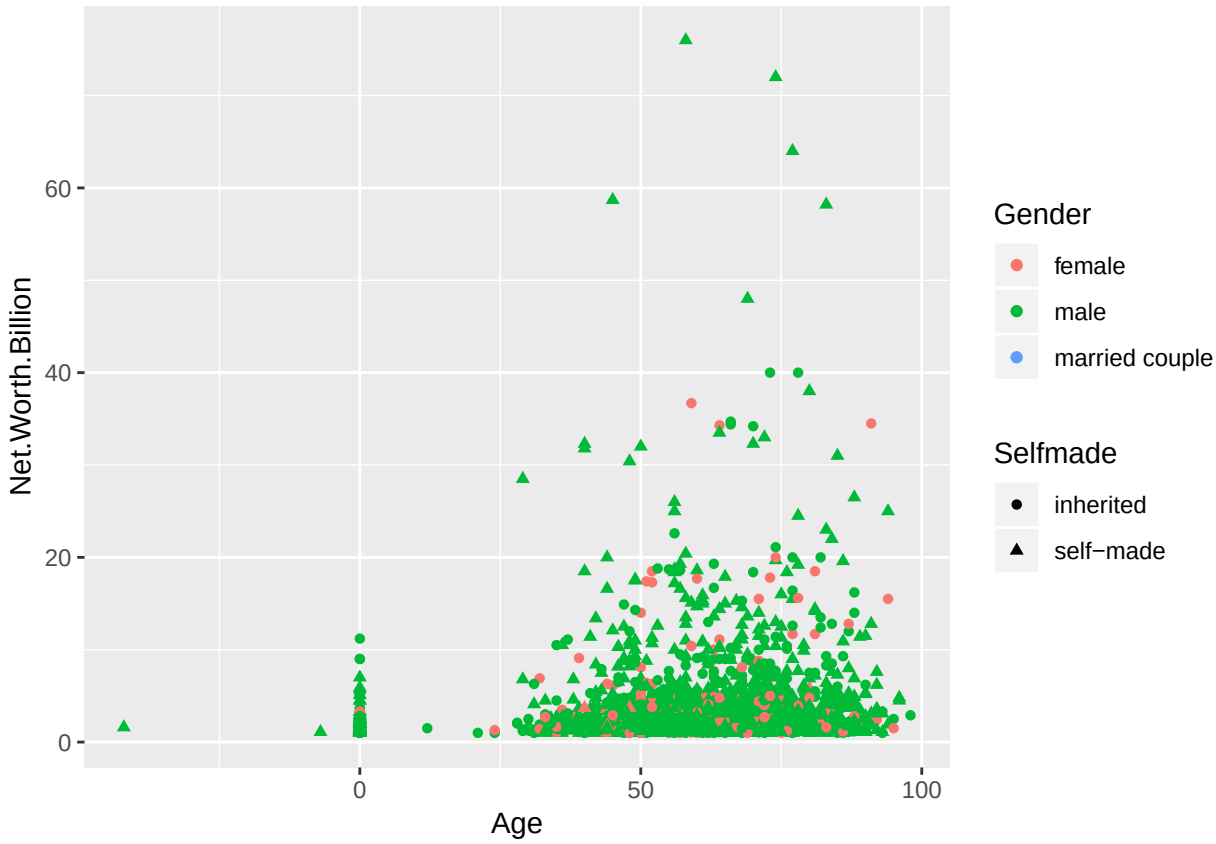
the columns that we deleted

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

Analysis

```
library(ggplot2)

# 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()
```

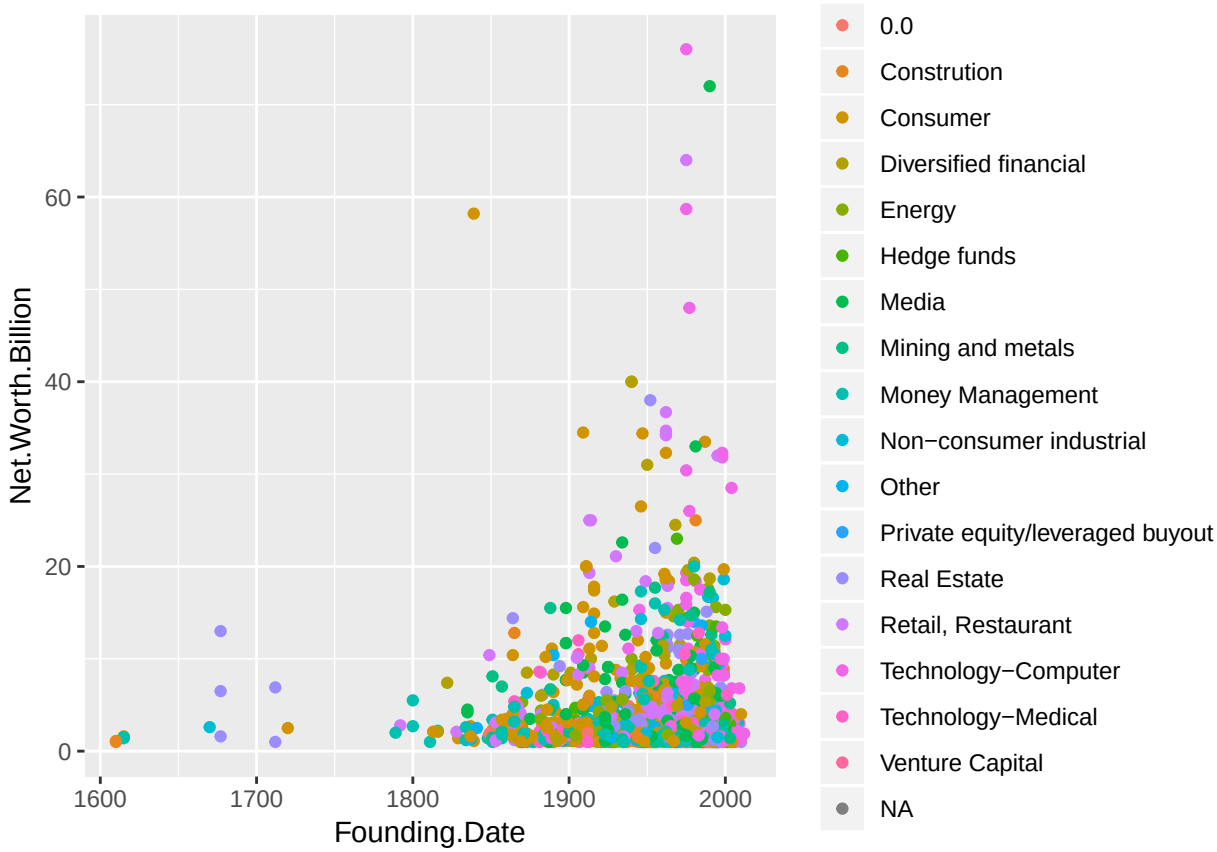


```r
# model of Age, Gender and Selfmade predicting the Net.Worth.Billion
mod_age_gen_self <- lm(Net.Worth.Billion ~ Age + Gender + Selfmade, gender_n_Selfmade)

summary(mod_age_gen_self)
```

```
##
## Call:
## lm(formula = Net.Worth.Billion ~ Age + Gender + Selfmade, data = gender_n_Selfmade)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.821 -2.195 -1.447 -0.002 72.519
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.296269   0.466815   4.919 9.29e-07 ***
## Age                     0.029061   0.005353   5.429 6.23e-08 ***
## Gendermale             -0.177852   0.388596  -0.458    0.647
## Gendermarried couple -0.796269   5.252586  -0.152    0.880
## Selfmadeself-made    -0.323180   0.238556  -1.355    0.176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.232 on 2375 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.01306,    Adjusted R-squared:  0.0114
## F-statistic: 7.858 on 4 and 2375 DF,  p-value: 2.738e-06
```

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



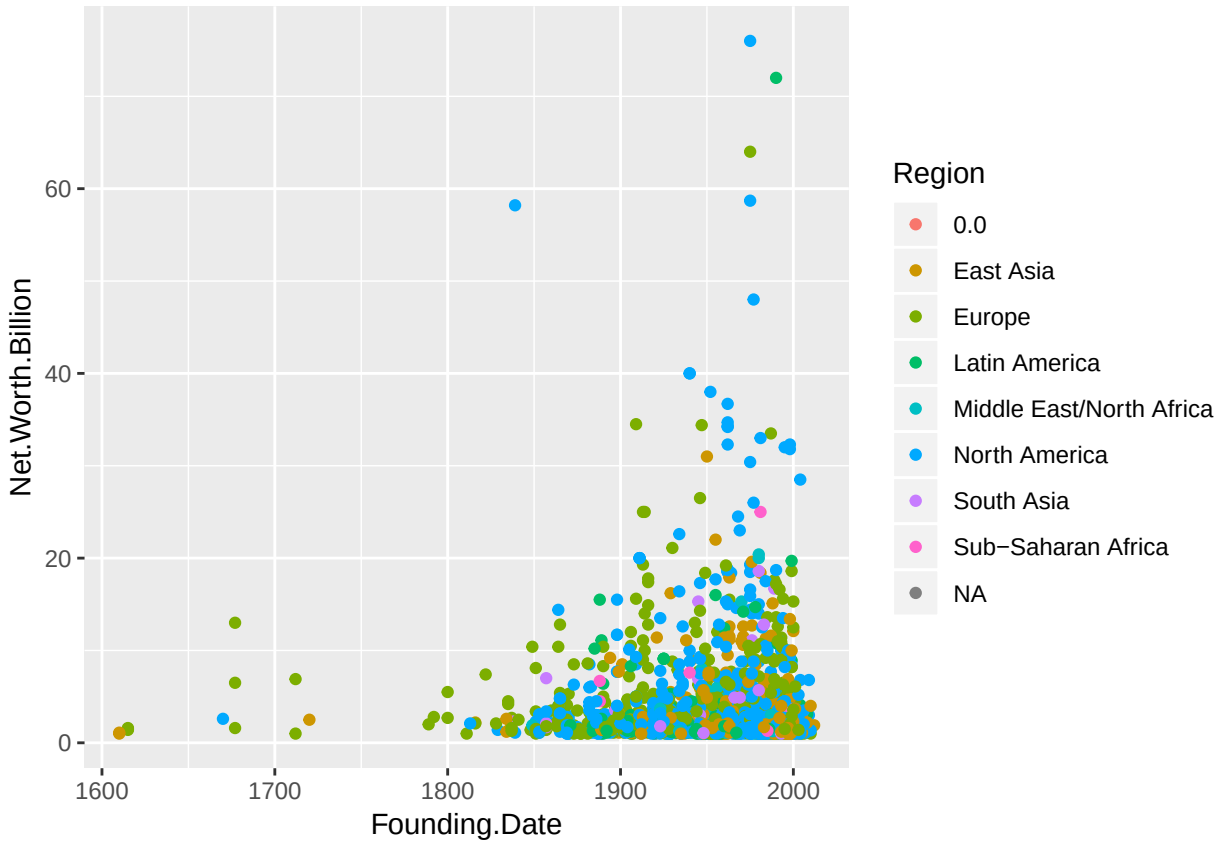

```
# model of Founding date and Industry predicting the Net.Worth.Billion
mod_date_ind <- lm(Net.Worth.Billion ~ Founding.Date + Industry, gender_n_Selfmade)

summary(mod_date_ind)
```

```
##
## Call:
## lm(formula = Net.Worth.Billion ~ Founding.Date + Industry, data = gender_n_Selfmade)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.123 -2.128 -1.272  0.026 71.077
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                          7.416251   5.041075   1.471
## Founding.Date                       -0.002946   0.002502  -1.177
## IndustryConstrution                  0.749214   1.463843   0.512
## IndustryConsumer                     2.069055   1.388273   1.490
## IndustryDiversified financial        2.564366   1.425578   1.799
## IndustryEnergy                       1.660249   1.439860   1.153
## IndustryHedge funds                  1.778960   1.508476   1.179
## IndustryMedia                        2.289904   1.412314   1.621
## IndustryMining and metals            1.503266   1.471726   1.021
## IndustryMoney Management             1.119188   1.405656   0.796
## IndustryNon-consumer industrial      1.708339   1.460217   1.170
## IndustryOther                        1.076110   1.480390   0.727
```

```
## IndustryPrivate equity/leveraged buyout  1.946519   1.711825   1.137
## IndustryReal Estate                      1.412992   1.404374   1.006
## IndustryRetail, Restaurant               2.541413   1.402521   1.812
## IndustryTechnology-Computer              3.325548   1.417553   2.346
## IndustryTechnology-Medical               1.021819   1.456835   0.701
## IndustryVenture Capital                  0.175507   2.269711   0.077
##                                         Pr(>|t|)
## (Intercept)                               0.1414
## Founding.Date                             0.2392
## IndustryConstrution                       0.6088
## IndustryConsumer                          0.1362
## IndustryDiversified financial             0.0722 .
## IndustryEnergy                            0.2490
## IndustryHedge funds                       0.2384
## IndustryMedia                             0.1051
## IndustryMining and metals                 0.3071
## IndustryMoney Management                  0.4260
## IndustryNon-consumer industrial           0.2421
## IndustryOther                             0.4673
## IndustryPrivate equity/leveraged buyout   0.2556
## IndustryReal Estate                       0.3144
## IndustryRetail, Restaurant                0.0701 .
## IndustryTechnology-Computer               0.0191 *
## IndustryTechnology-Medical                0.4831
## IndustryVenture Capital                   0.9384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.117 on 2524 degrees of freedom
##   (21 observations deleted due to missingness)
## Multiple R-squared:  0.018,  Adjusted R-squared:  0.01138
## F-statistic: 2.721 on 17 and 2524 DF,  p-value: 0.0001739
```

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



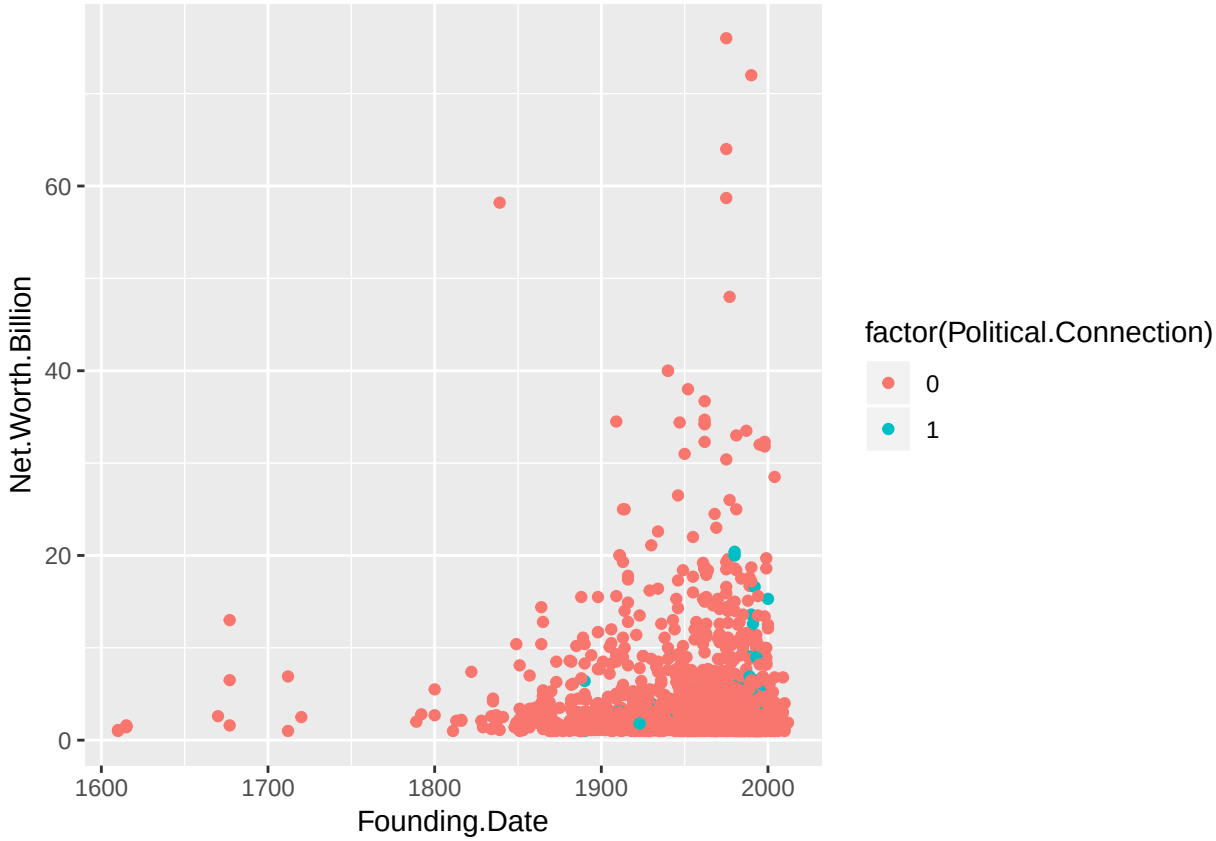

```
# model of Founding date and region predicting the Net.Worth.Billion
mod_date_reg <- lm(Net.Worth.Billion ~ Founding.Date + Region, gender_n_Selfmade)

summary(mod_date_reg)
```

```
##
## Call:
## lm(formula = Net.Worth.Billion ~ Founding.Date + Region, data = gender_n_Selfmade)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.912 -2.132 -1.410 -0.065 72.130
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.7459789  6.9675704   0.251    0.802
## Founding.Date                   0.0001845  0.0024534   0.075    0.940
## RegionEast Asia                 0.8314764  5.1427431   0.162    0.872
## RegionEurope                    1.7289597  5.1405555   0.336    0.737
## RegionLatin America             1.1236231  5.1511991   0.218    0.827
## RegionMiddle East/North Africa 0.6280003  5.1595670   0.122    0.903
## RegionNorth America             1.7599649  5.1396791   0.342    0.732
## RegionSouth Asia                1.0903881  5.1743112   0.211    0.833
## RegionSub-Saharan Africa        1.7968460  5.2642379   0.341    0.733
##
## Residual standard error: 5.136 on 2534 degrees of freedom
##   (20 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.006673,   Adjusted R-squared:  0.003537
## F-statistic: 2.128 on 8 and 2534 DF,  p-value: 0.03025
```

```
#color code by political connection
ggplot(gender_n_Selfmade, aes(x = Founding.Date, y = Net.Worth.Billion, color = factor(Political.Connec
  geom_point()
```




```
# model of Founding date and Political.Connection predicting the Net.Worth.Billion



mod_date_poli <- lm(Net.Worth.Billion ~ Founding.Date + factor(Political.Connection), gender_n_Selfmade

summary(mod_date_poli)
```

```
##
## Call:
## lm(formula = Net.Worth.Billion ~ Founding.Date + factor(Political.Connection),
##     data = gender_n_Selfmade)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.097 -2.202 -1.515 -0.113 72.477
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.784245   4.732809   1.222    0.222
## Founding.Date                 -0.001145   0.002421  -0.473    0.636
```

```
## factor(Political.Connection)1  0.477188   0.609279   0.783    0.434
##
## Residual standard error: 5.147 on 2540 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.0003074,  Adjusted R-squared:  -0.0004797
## F-statistic: 0.3906 on 2 and 2540 DF,  p-value: 0.6767
```

```
# color code by the type of company
ggplot(gender_n_Selfmade, aes(x = Age, y = Net.Worth.Billion, color = Company.type)) +
  geom_point()
```

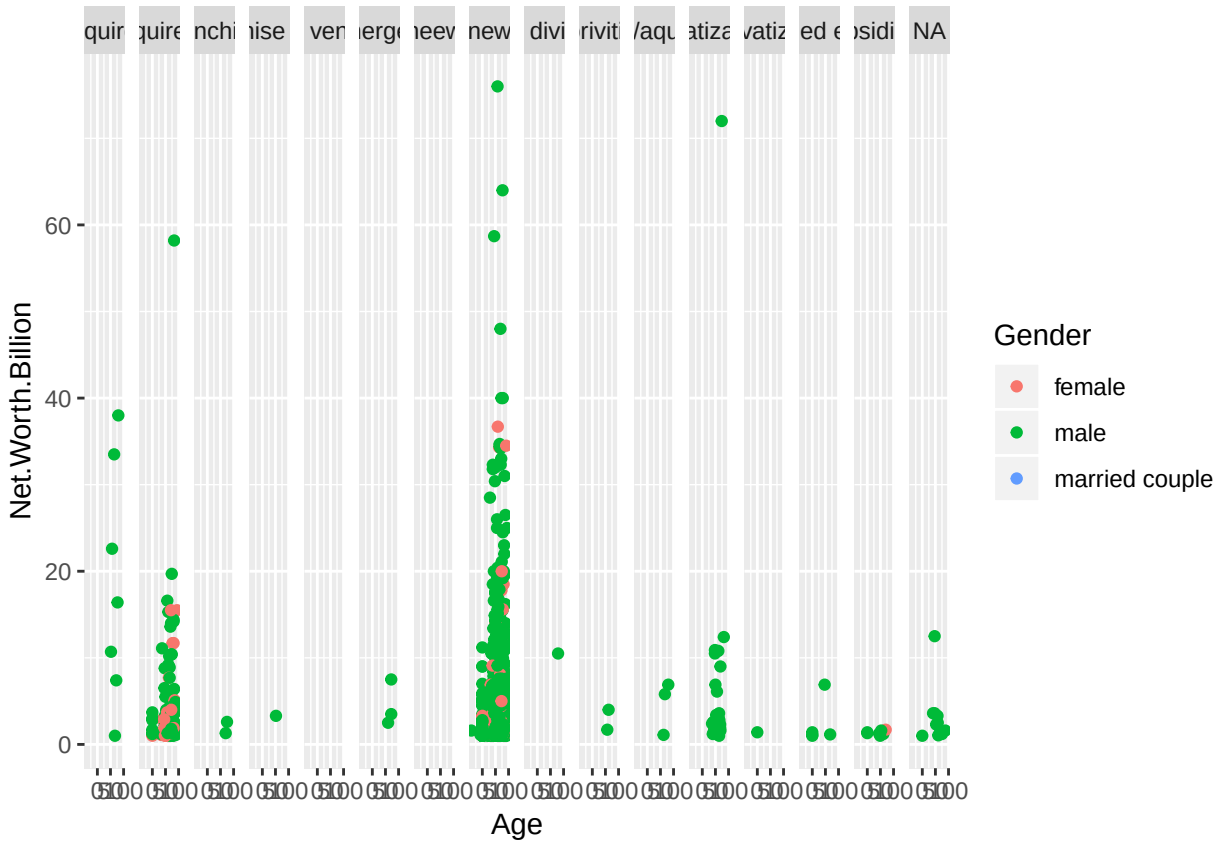

```
# model of Founding date and Industry predicting the Net.Worth.Billion
mod_age_comp <- lm(Net.Worth.Billion ~ Age + Company.type, gender_n_Selfmade)

summary(mod_age_comp)
```

```
##
## Call:
## lm(formula = Net.Worth.Billion ~ Age + Company.type, data = gender_n_Selfmade)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.518  -2.193  -1.386   0.051  72.454
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     16.674433   1.991757   8.372  < 2e-16
## Age                              0.027519   0.005338   5.155 2.74e-07
```

```
## Company.typeaquired                -14.652471   1.996849  -7.338 2.97e-13
## Company.typefranchise              -16.719572   4.156887  -4.022 5.95e-05
## Company.typefranchise rights       -14.805430   5.542938  -2.671 0.007614
## Company.typemerger                 -14.045736   3.577590  -3.926 8.88e-05
## Company.typenew                    -14.724944   1.963299  -7.500 8.98e-14
## Company.typenew division            -8.348447   5.542749  -1.506 0.132152
## Company.typenew, privitization     -15.461823   4.156963  -3.719 0.000204
## Company.typenew/aquired            -13.991601   3.577616  -3.911 9.46e-05
## Company.typeprivatization          -13.308594   2.117378  -6.285 3.88e-10
## Company.typeprivatized             -15.274433   5.553848  -2.750 0.006001
## Company.typestate owned enterprise -14.997374   3.044695  -4.926 8.99e-07
## Company.typesubsidiary             -16.473988   2.686467  -6.132 1.01e-09
##
## (Intercept)                        ***
## Age                                ***
## Company.typeaquired                ***
## Company.typefranchise              ***
## Company.typefranchise rights       **
## Company.typemerger                 ***
## Company.typenew                    ***
## Company.typenew division
## Company.typenew, privitization     ***
## Company.typenew/aquired            ***
## Company.typeprivatization          ***
## Company.typeprivatized             **
## Company.typestate owned enterprise ***
## Company.typesubsidiary             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.184 on 2355 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.03741,    Adjusted R-squared:  0.0321
## F-statistic: 7.041 on 13 and 2355 DF,  p-value: 1.392e-13
```

```r
# Separate by each company type
ggplot(gender_n_Selfmade, aes(x = Age, y = Net.Worth.Billion, color = Gender)) +
  geom_point() +
  facet_grid(. ~Company.type)
```


quir
uire
nchi
nise
ver
erg
neev
new
divi
rivit
/aqu
atiza
vatiz
ed e
sidi
NA
60
40
20
0
Net.Worth.Billion
Age
Gender
female
male
married couple

Analysis: