SAP - projekt - Milijarderi

Uspjeh učenika u nastavi

Dora Bezuk, Marcela Matas, Josip Arelic, Domagoj Marinello

13.11.2022.

Uvod

Pitanja:

1. Ima li neki kontinent statistički značajno više miljarda?
2. Jesu li milijarderi koji su nasljedili bogastvo statistički značajno bogatiji od onih koji nisu?
3. Možete li iz danih varijabli predvidjeti njihovo bogatstvo?
4. Kada biste birali karijeru isključivo prema kriteriju da se obogatite, koju biste industriju izabrali?

Dodatna pitanja:

5. ???

Deskriptivna analiza

```
# Pomoćna funkcija za izbacivanje stršećih vrijednosti
remove_outliers <- function(data, data_column) {
  quartiles <- quantile(data_column, probs=c(.25, .75), na.rm = FALSE)
  IQR <- IQR(data_column)
  Lower <- quartiles[1] - 1.5*IQR
  Upper <- quartiles[2] + 1.5*IQR

  return(subset(data, data_column >= Lower & data_column <= Upper))
}

cat('\n Dimenzija podataka: ', dim(bill_data))
```

```
##
##  Dimenzija podataka:  2614 22
```

```
for (col_name in names(bill_data)){
  if (sum(is.na(bill_data[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu'
        ,col_name, ': ', sum(is.na(bill_data[,col_name])),'\n')
  }
}
```

```
## Ukupno nedostajućih vrijednosti za varijablu company.name :  38
## Ukupno nedostajućih vrijednosti za varijablu company.relationship :  46
## Ukupno nedostajućih vrijednosti za varijablu company.sector :  23
```

```
## Ukupno nedostajućih vrijednosti za varijablu company.type :  36
## Ukupno nedostajućih vrijednosti za varijablu demographics.gender :  34
## Ukupno nedostajućih vrijednosti za varijablu wealth.type :  22
## Ukupno nedostajućih vrijednosti za varijablu wealth.how.category :  1
## Ukupno nedostajućih vrijednosti za varijablu wealth.how.industry :  1
```

Postoje podaci koji nedostaju. Što s njima?

Pitanja

1. Ima li neki kontinent statistički značajno više miljarda?

2. Jesu li milijarderi koji su nasljedili bogastvo statistički značajno bogatiji od onih koji nisu?

```
# Učitavanje podataka iz excel datoteke

inherited = bill_data[bill_data$wealth.how.inherited!="not inherited",]
print(inherited)
```

```
## # A tibble: 926 x 22
##    name            rank  year company.founded company.name     company.relation~
##    <chr>          <dbl> <dbl>           <dbl> <chr>            <chr>
##  1 Oeri Hoffman ~     3  1996            1896 F. Hoffmann-La ~ <NA>
##  2 Walter Thomas~     6  1996            1963 Sun Hung Kai Pr~ Relation
##  3 Charles Koch       6  2014            1940 Koch industries  relation
##  4 David Koch         6  2014            1940 Koch industries  relation
##  5 Jim Walton         7  2001            1962 Walmart          relation
##  6 Yoshiaki Tsut~     8  1996            1894 Seibu Corporati~ relation
##  7 John Walton        8  2001            1962 Walmart          relation
##  8 Theo and Karl~     9  1996            1913 Aldi Nord        Relation
##  9 S Robson Walt~     9  2001            1962 Walmart          relation
## 10 Christy Walton     9  2014            1962 Walmart          relation
## # ... with 916 more rows, and 16 more variables: company.sector <chr>,
## #   company.type <chr>, demographics.age <dbl>, demographics.gender <chr>,
## #   location.citizenship <chr>, location.country code <chr>,
## #   location.gdp <dbl>, location.region <chr>, wealth.type <chr>,
## #   wealth.worth in billions <dbl>, wealth.how.category <chr>,
## #   wealth.how.from emerging <chr>, wealth.how.industry <chr>,
## #   wealth.how.inherited <chr>, wealth.how.was founder <chr>, ...
```

```
non_inherited = bill_data[bill_data$wealth.how.inherited=="not inherited",]
print(non_inherited)
```

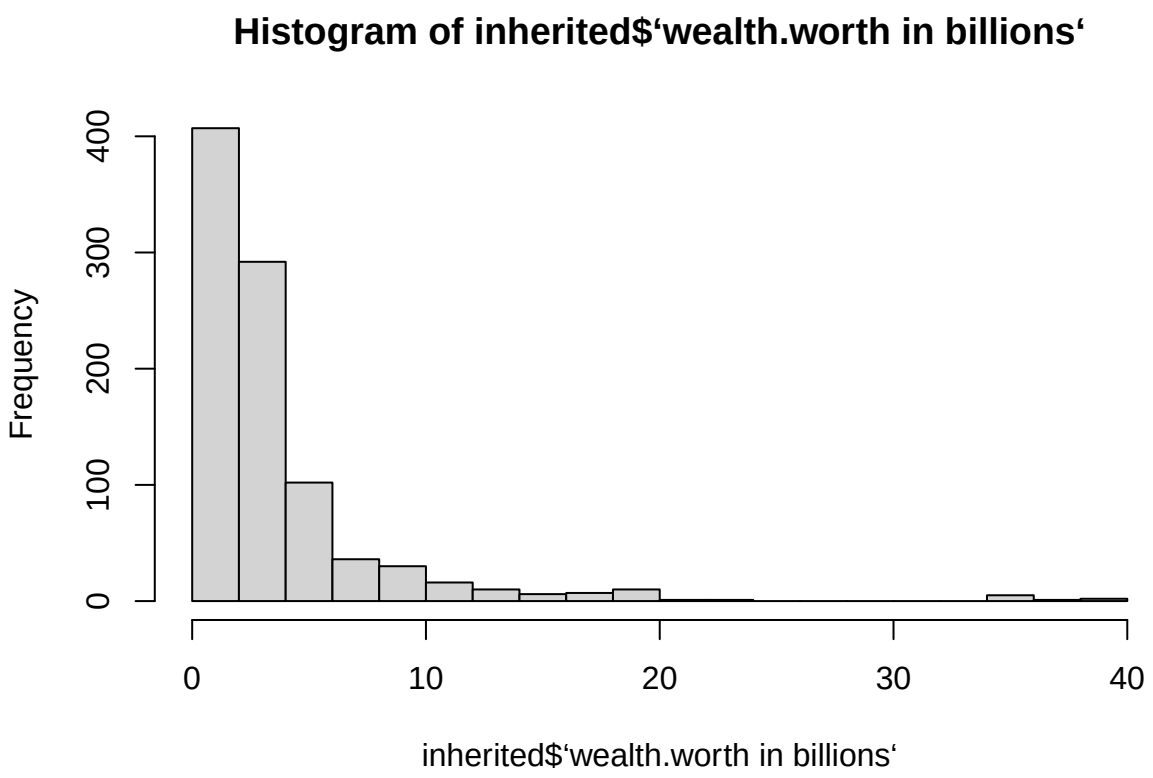
```
## # A tibble: 1,688 x 22
##    name             rank  year company.founded company.name    company.relatio~
##    <chr>           <dbl> <dbl>           <dbl> <chr>           <chr>
##  1 Bill Gates          1  1996            1975 Microsoft       founder
##  2 Bill Gates          1  2001            1975 Microsoft       founder
##  3 Bill Gates          1  2014            1975 Microsoft       founder
##  4 Warren Buffett      2  1996            1962 Berkshire Hath~ founder
##  5 Warren Buffett      2  2001            1962 Berkshire Hath~ founder
##  6 Carlos Slim Helu    2  2014            1990 Telmex          founder
##  7 Paul Allen          3  2001            1975 Microsoft       founder
##  8 Amancio Ortega      3  2014            1975 Zara            founder
```

```
##  9 Lee Shau Kee          4  1996            1976 Henderson Land~ founder/chairman
## 10 Larry Ellison         4  2001            1977 Oracle          founder
## # ... with 1,678 more rows, and 16 more variables: company.sector <chr>,
## #   company.type <chr>, demographics.age <dbl>, demographics.gender <chr>,
## #   location.citizenship <chr>, location.country code <chr>,
## #   location.gdp <dbl>, location.region <chr>, wealth.type <chr>,
## #   wealth.worth in billions <dbl>, wealth.how.category <chr>,
## #   wealth.how.from emerging <chr>, wealth.how.industry <chr>,
## #   wealth.how.inherited <chr>, wealth.how.was founder <chr>, ...
```

```
inherited_mean = mean(inherited$`wealth.worth in billions`)
print(inherited_mean)
```

```
## [1] 3.750756
```

```
hist(inherited$`wealth.worth in billions`, breaks = 20)
```

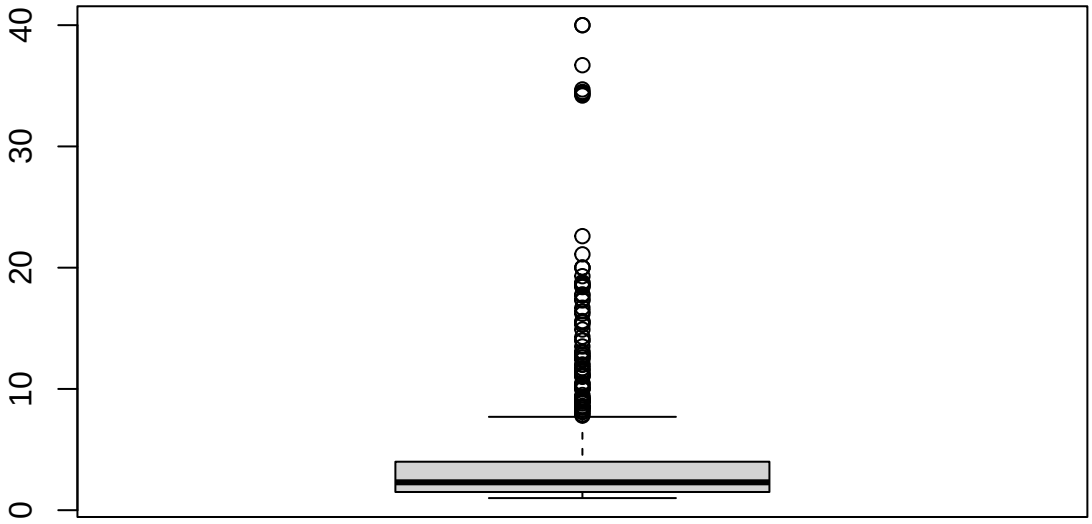



```
non_inherited_mean = mean(non_inherited$`wealth.worth in billions`)
print(non_inherited_mean)
```

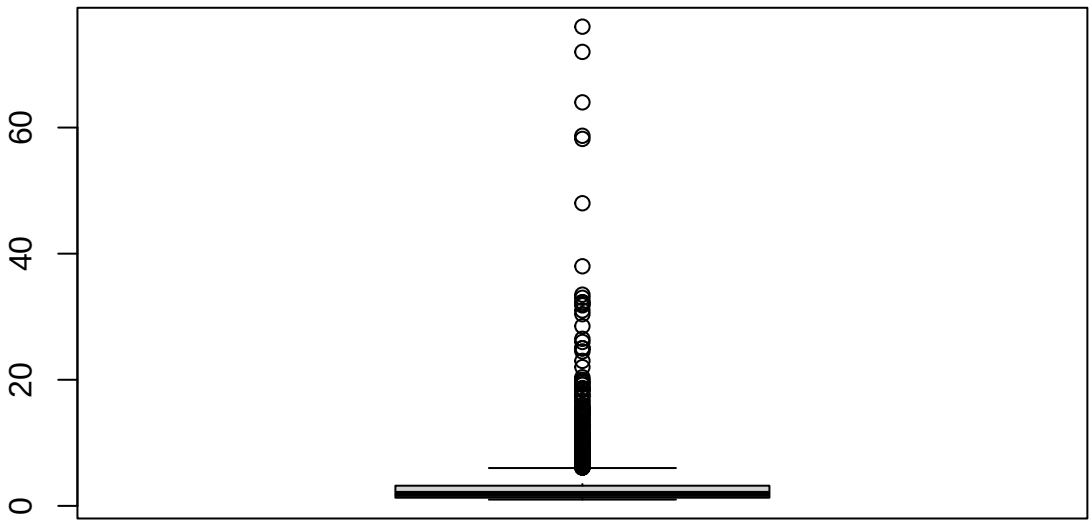
```
## [1] 3.411908
```

```
hist(non_inherited$`wealth.worth in billions`, breaks = 20)
```


```
boxplot(inherited$`wealth.worth in billions`)
```




```
boxplot(non_inherited$`wealth.worth in billions`)
```




```
wilcox.test(inherited_mean, non_inherited_mean, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  inherited_mean and non_inherited_mean
## W = 1, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
# jel smijem odrezat dio podataka (pogledati histogram i box plot), koji dio?
# ako ne, koji test koristi? t-test za mediane umjesto meana? ili neparametarski test? kao sto je wilco
```

Formiranje hipoteza

Vizualizacija podataka

Pretpostavke za provođenje testa

Test xy

Provođenje T-testa

Zaključak

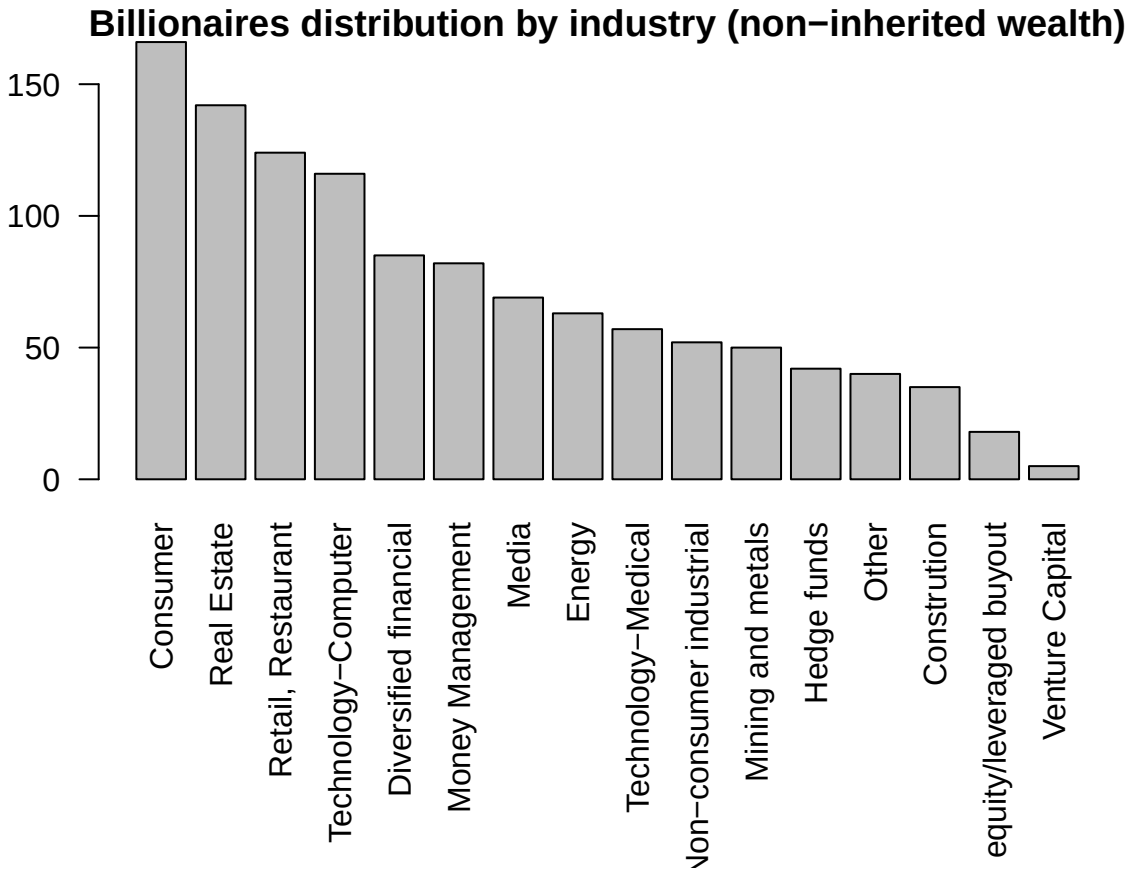
3. Možete li iz danih varijabli predvidjeti njihovo bogatstvo?

4. Kada biste birali karijeru isključivo prema kriteriju da se obogatite, koju biste industriju izabrali?

Pretpostavljamo da karijerom u određenoj industriji, a ne nasljedstvom zarađujemo novac. Zbog toga gledamo samo milijardere koji nisu nasljedili svoje bogatstvo. Također, zanimaju nas samo najnoviji milijarderi odnosno oni s popisa iz 2014. godine.

```
#
non_inherited_2014 <- non_inherited[non_inherited$year == 2014,]

par(mar=c(10,5,1,1))
barplot(sort(table(subset(non_inherited_2014$wealth.how.industry, non_inherited_2014$wealth.how.industry
        main = "Billionaires distribution by industry (non-inherited wealth)",
        las = 2)
```


Billionaires distribution by industry (non−inherited wealth)
150
100
50
0
Consumer
Real Estate
Retail, Restaurant
Technology−Computer
Diversified financial
Money Management
Media
Energy
Technology−Medical
Non−consumer industrial
Mining and metals
Hedge funds
Other
Construtions
equity/leveraged buyout
Venture Capital