

# 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 milijarda?
2. Jesu li milijarderi koji su naslijedili 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 milijarda?
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 compa~1 compa~2 compa~3 compa~4 compa~5 demog~6 demog~7
##   <chr>    <dbl> <dbl>   <dbl> <chr>   <chr>   <chr>   <chr>   <dbl> <chr>
## 1 Oeri Hof~    3  1996   1896 F. Hof~ <NA>   pharma~ new      0 <NA>
## 2 Walter T~    6  1996   1963 Sun Hu~ Relati~ real e~ new      0 male
## 3 Charles ~    6  2014   1940 Koch i~ relati~ Oil re~ new     78 male
## 4 David Ko~    6  2014   1940 Koch i~ relati~ Oil re~ new     73 male
## 5 Jim Walt~    7  2001   1962 Walmart relati~ retail new     53 male
## 6 Yoshiaki~    8  1996   1894 Seibu ~ relati~ real e~ aquired 61 male
## 7 John Wal~    8  2001   1962 Walmart relati~ retail new     55 male
## 8 Theo and~    9  1996   1913 Aldi N~ Relati~ grocer~ new      0 male
## 9 S Robson~    9  2001   1962 Walmart relati~ retail new     57 male
## 10 Christy ~    9  2014   1962 Walmart relati~ retail new     59 female
## # ... with 916 more rows, 12 more variables: 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>, `wealth.how.was political` <chr>, and
## #   abbreviated variable names 1: company.founded, 2: company.name, ...
```

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

```
## # A tibble: 1,688 x 22
##   name      rank  year compa~1 compa~2 compa~3 compa~4 compa~5 demog~6 demog~7
##   <chr>    <dbl> <dbl>   <dbl> <chr>   <chr>   <chr>   <chr>   <dbl> <chr>
## 1 Bill Gat~    1  1996   1975 Micros~ founder Softwa~ new     40 male
## 2 Bill Gat~    1  2001   1975 Micros~ founder Softwa~ new     45 male
## 3 Bill Gat~    1  2014   1975 Micros~ founder Softwa~ new     58 male
## 4 Warren B~    2  1996   1962 Berksh~ founder Finance new     65 male
## 5 Warren B~    2  2001   1962 Berksh~ founder Finance new     70 male
## 6 Carlos S~    2  2014   1990 Telmex founder Commun~ privat~ 74 male
## 7 Paul All~    3  2001   1975 Micros~ founder techno~ new     48 male
## 8 Amancio ~    3  2014   1975 Zara   founder Fashion new     77 male
```

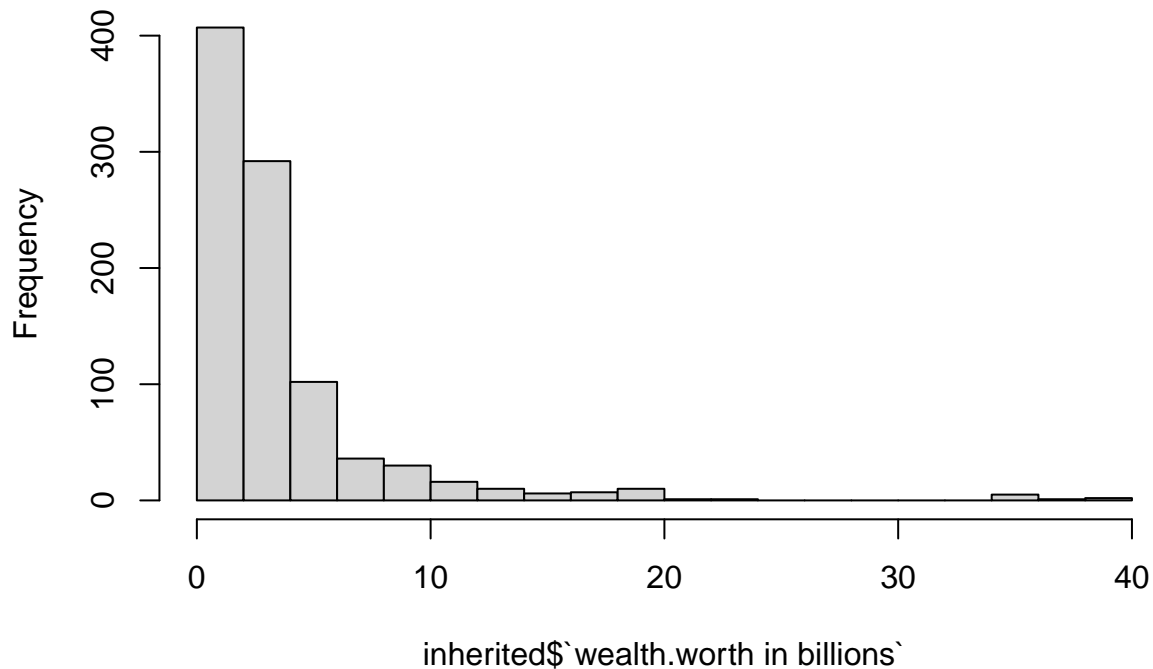
```
## 9 Lee Chau~      4 1996      1976 Hender~ founde~ real e~ new      68 male
## 10 Larry El~     4 2001      1977 Oracle founder softwa~ new      56 male
## # ... with 1,678 more rows, 12 more variables: location.citizenship <chr>,
## #   `location.country code` <chr>, location.gdp <dbl>, location.region <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>, `wealth.how.was political` <chr>, and
## #   abbreviated variable names 1: company.founded, 2: company.name, ...
```

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

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

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

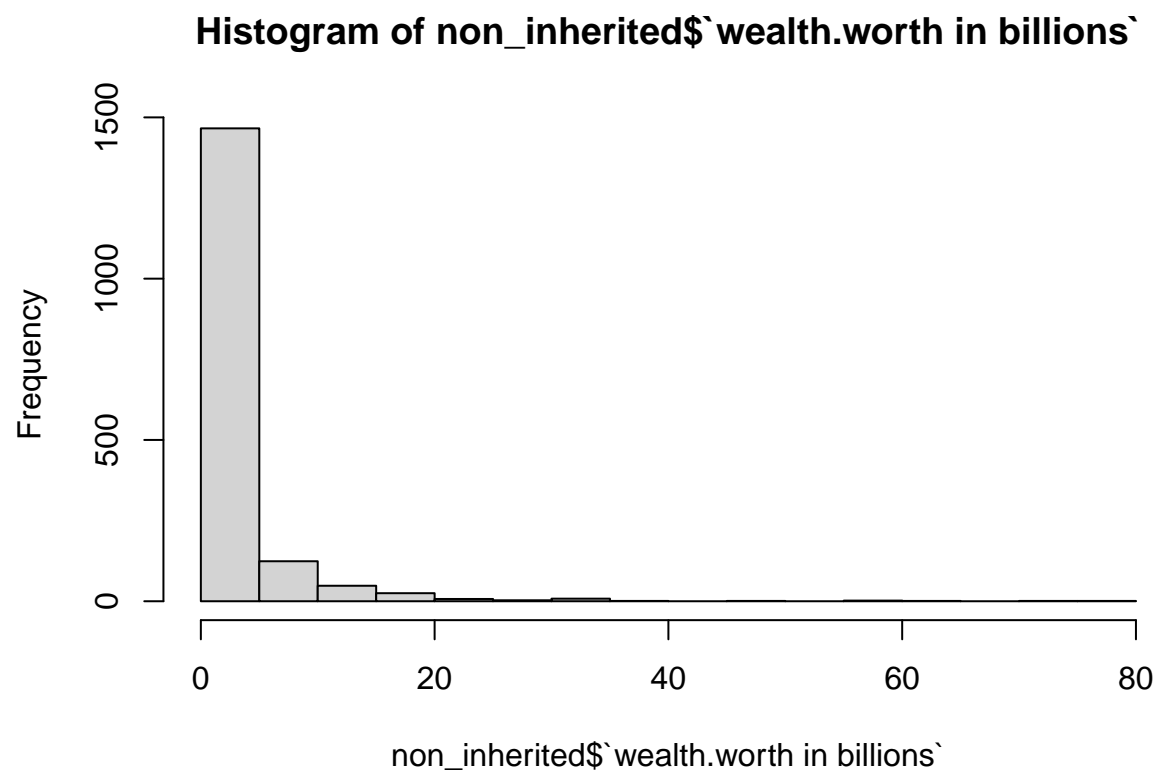
### Histogram of inherited\$`wealth.worth in billions`



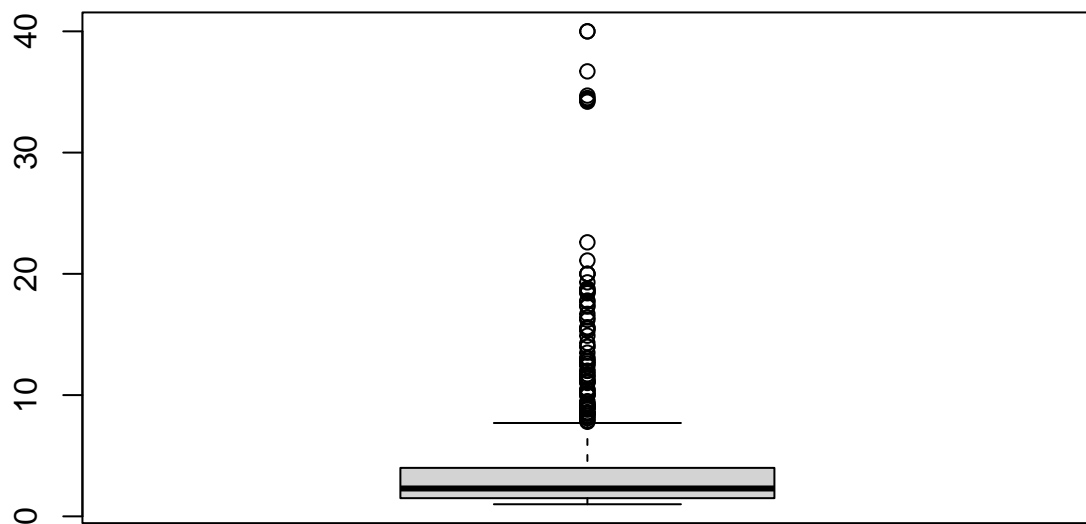
```
non_inherited_mean = mean(non_inherited$`wealth.worth in billions`)
print(non_inherited_mean)
```

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

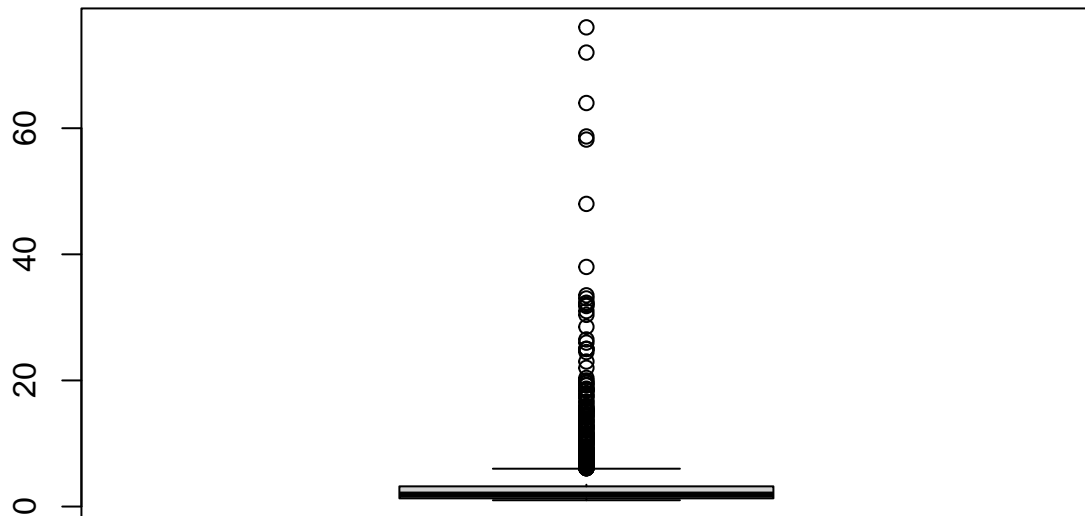
```
hist(non_inherited$`wealth.worth in billions`, breaks = 20)
```



```
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)
```

