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?

```
summary(bill_data)
```

```
##      name                rank            year       company.founded
##  Length:2614        Min.   :   1.0   Min.   :1996   Min.   :   0
##  Class :character   1st Qu.: 215.0   1st Qu.:2001   1st Qu.:1936
##  Mode  :character   Median : 430.0   Median :2014   Median :1963
##                     Mean   : 599.7   Mean   :2008   Mean   :1925
##                     3rd Qu.: 988.0   3rd Qu.:2014   3rd Qu.:1985
##                     Max.   :1565.0   Max.   :2014   Max.   :2012
##  company.name       company.relationship company.sector     company.type
##  Length:2614        Length:2614          Length:2614        Length:2614
##  Class :character   Class :character     Class :character   Class :character
##  Mode  :character   Mode  :character     Mode  :character   Mode  :character
##
##
##
##  demographics.age demographics.gender location.citizenship
##  Min.   :-42.00   Length:2614         Length:2614
##  1st Qu.: 47.00   Class :character    Class :character
##  Median : 59.00   Mode  :character    Mode  :character
##  Mean   : 53.34
##  3rd Qu.: 70.00
##  Max.   : 98.00
##  location.country code  location.gdp      location.region
##  Length:2614          Min.   :0.000e+00   Length:2614
##  Class :character     1st Qu.:0.000e+00   Class :character
##  Mode  :character     Median :0.000e+00   Mode  :character
##                       Mean   :1.769e+12
##                       3rd Qu.:7.250e+11
##                       Max.   :1.060e+13
##  wealth.type        wealth.worth in billions wealth.how.category
##  Length:2614        Min.   : 1.000           Length:2614
##  Class :character   1st Qu.: 1.400           Class :character
##  Mode  :character   Median : 2.000           Mode  :character
##                     Mean   : 3.532
##                     3rd Qu.: 3.500
##                     Max.   :76.000
##  wealth.how.from emerging wealth.how.industry wealth.how.inherited
##  Length:2614              Length:2614         Length:2614
##  Class :character         Class :character    Class :character
##  Mode  :character         Mode  :character    Mode  :character
##
##
##
##  wealth.how.was founder wealth.how.was political
##  Length:2614            Length:2614
##  Class :character       Class :character
```

```
## Mode  :character      Mode  :character
##
##
##
```

```
sapply(bill_data, class)
```

```
##                   name                   rank                   year
##             "character"              "numeric"              "numeric"
##         company.founded           company.name   company.relationship
##              "numeric"            "character"            "character"
##         company.sector           company.type         demographics.age
##            "character"            "character"              "numeric"
##     demographics.gender    location.citizenship   location.country code
##            "character"            "character"            "character"
##           location.gdp         location.region            wealth.type
##              "numeric"            "character"            "character"
## wealth.worth in billions    wealth.how.category wealth.how.from emerging
##              "numeric"            "character"            "character"
##     wealth.how.industry    wealth.how.inherited    wealth.how.was founder
##            "character"            "character"            "character"
## wealth.how.was political
##            "character"
```
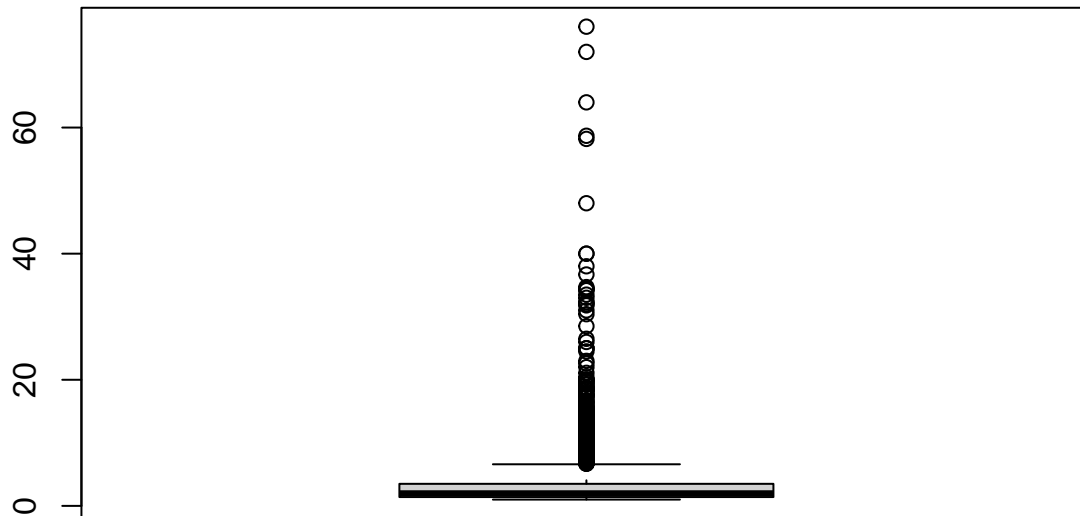
Naš dataset sastoji se od character i numeric varijabli.

Prvo promotrimo numeričke varijable.

```
hist(bill_data$`wealth.worth in billions` ,main='wealth worth in billions', xlab='wealth', ylab='Frequen
```

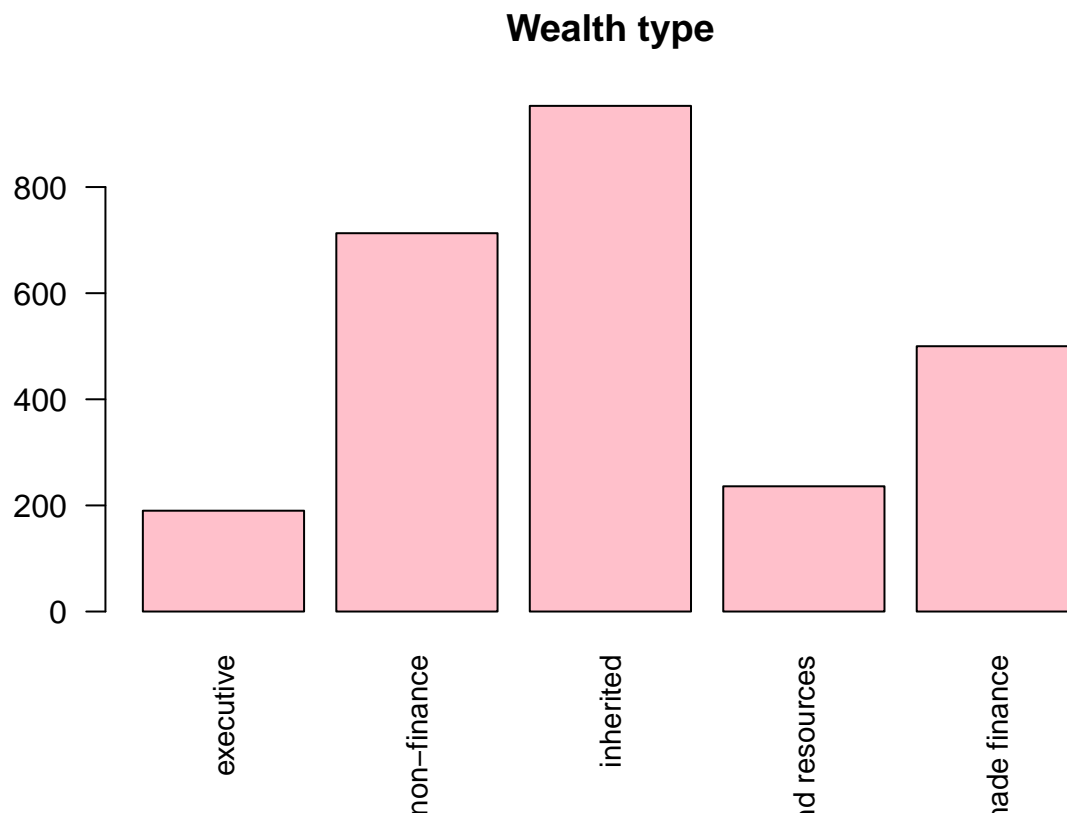**wealth worth in billions**
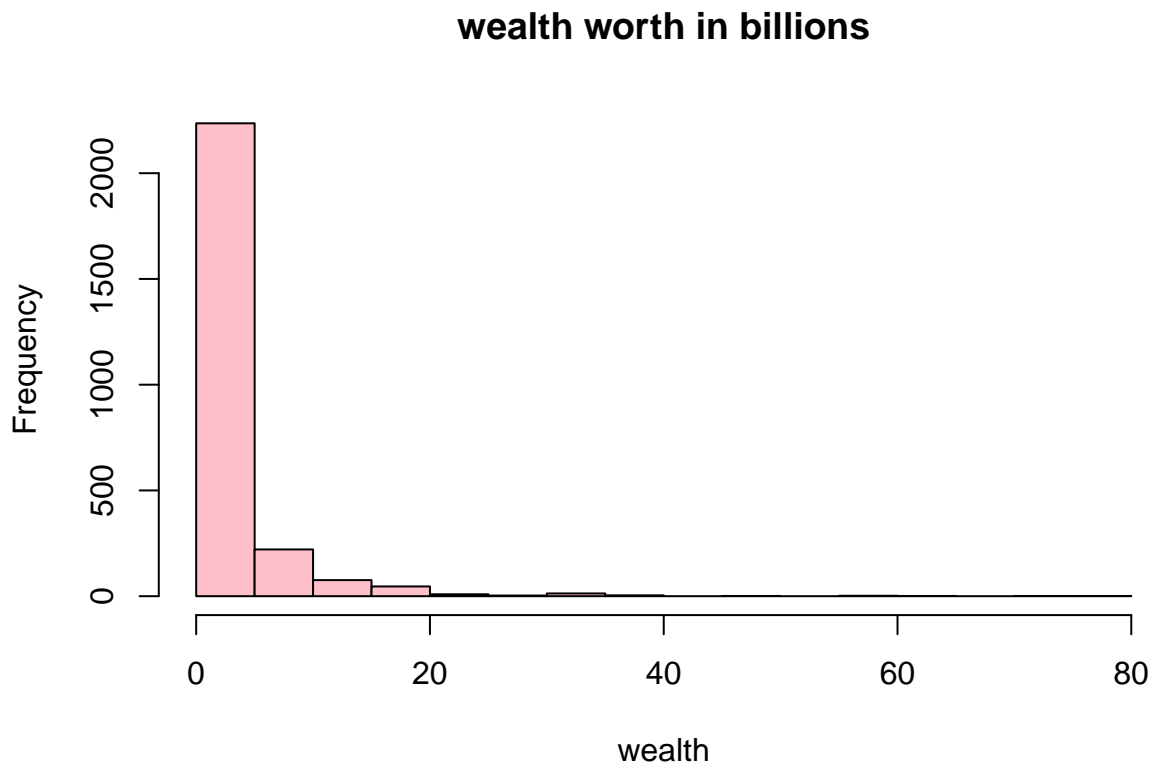

```
boxplot(bill_data$`wealth.worth in billions`)
```
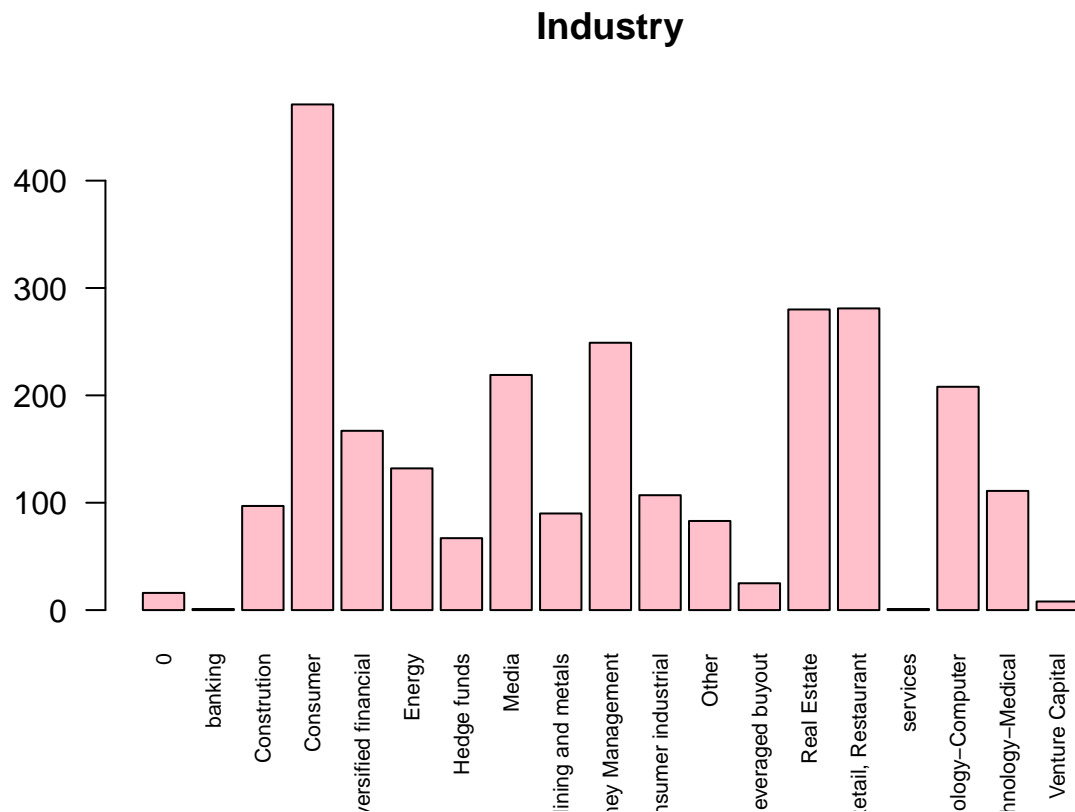
```
summary(bill_data$`wealth.worth in billions`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.400   2.000   3.532   3.500  76.000
```

```
barplot(table(bill_data$wealth.type),las=2,cex.names=.9,main='Wealth type',col="pink")
```

Wealth type


```
barplot(table(bill_data$wealth.how.industry),las=2,cex.names=.7,main='Industry',col="pink")
```

Industry


```
print('Podjela po spolu: ')
```

```
## [1] "Podjela po spolu: "
```

```
table(bill_data$demographics.gender)
```

```
##
##        female          male married couple
##           249          2328               3
```

Pitanja

1. Ima li neki kontinent statistički značajno više miljarda?

```
levels(factor(bill_data$location.region))
```

```
## [1] "0"                        "East Asia"
## [3] "Europe"                   "Latin America"
## [5] "Middle East/North Africa" "North America"
## [7] "South Asia"               "Sub-Saharan Africa"
```

```
class(bill_data$location.region)
```

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

Treba li tip stupca biti factor?

Ima li nedostajućih vrijednosti?

```
# is.na ce nam vratiti logical vektor koji ima TRUE na mjestima gdje ima NA:
sum(is.na(bill_data$location.region))
```

```
## [1] 0
```

Nema nedostajućih vrijednosti

```
table(bill_data$location.region)
```

```
##
##                          0              East Asia                  Europe
##                          1                    535                     698
##              Latin America Middle East/North Africa           North America
##                        182                    117                     992
##                 South Asia     Sub-Saharan Africa
##                         69                     20
```

```
bill_data$location.citizenship[bill_data$location.region == "Middle East/North Africa"]
```

```
##   [1] "Saudi Arabia"         "Saudi Arabia"         "Saudi Arabia"
##   [4] "Saudi Arabia"         "Kuwait"               "Turkey"
##   [7] "Saudi Arabia"         "Turkey"               "Kuwait"
##  [10] "Saudi Arabia"         "Turkey"               "Israel"
##  [13] "Turkey"               "Lebanon"              "Saudi Arabia"
##  [16] "Saudi Arabia"         "Lebanon"              "Saudi Arabia"
##  [19] "Saudi Arabia"         "Turkey"               "Israel"
##  [22] "Israel"               "Saudi Arabia"         "Israel"
##  [25] "Lebanon"              "Turkey"               "Israel"
##  [28] "United Arab Emirates" "Saudi Arabia"         "Saudi Arabia"
##  [31] "Israel"               "Turkey"               "United Arab Emirates"
##  [34] "Israel"               "Turkey"               "Israel"
##  [37] "Israel"               "United Arab Emirates" "Saudi Arabia"
##  [40] "Israel"               "Israel"               "Bahrain"
##  [43] "Saudi Arabia"         "Israel"               "Israel"
##  [46] "Saudi Arabia"         "Saudi Arabia"         "Turkey"
##  [49] "Saudi Arabia"         "Turkey"               "Israel"
##  [52] "Egypt"                "Algeria"              "Egypt"
##  [55] "Saudi Arabia"         "Lebanon"              "Lebanon"
##  [58] "Israel"               "Turkey"               "Turkey"
##  [61] "Egypt"                "Morocco"              "United Arab Emirates"
##  [64] "United Arab Emirates" "Israel"               "Israel"
##  [67] "Saudi Arabia"         "Egypt"                "Saudi Arabia"
##  [70] "Egypt"                "Lebanon"              "Turkey"
##  [73] "Turkey"               "Turkey"               "Morocco"
##  [76] "Egypt"                "Saudi Arabia"         "Turkey"
##  [79] "Turkey"               "Israel"               "Egypt"
##  [82] "Israel"               "Turkey"               "Turkey"
##  [85] "Turkey"               "Turkey"               "Turkey"
##  [88] "Turkey"               "Turkey"               "Lebanon"
##  [91] "Morocco"              "Turkey"               "Israel"
##  [94] "Israel"               "Kuwait"               "Kuwait"
##  [97] "Israel"               "Kuwait"               "Turkey"
## [100] "Turkey"               "Egypt"                "Israel"
## [103] "Morocco"              "Kuwait"               "Kuwait"
## [106] "Turkey"               "Lebanon"              "Lebanon"
## [109] "Oman"                 "Israel"               "Turkey"
## [112] "Turkey"               "Oman"                 "Turkey"
## [115] "Israel"               "Israel"               "Turkey"
```

Sada možemo združiti podatke ovisno o kontinentu.

Kopirajmo najprije podatke u novi data.frame kako ne bi promijenili prave vrijednosti.

```
bill_data_copy = data.frame(bill_data)
tracemem(bill_data)==tracemem(bill_data_copy)
```

[1] FALSE

```
untracemem(bill_data_copy)
untracemem(bill_data_copy)
```

```
# Zdruzimo Europu
for (column_name in c("Europe")){
  bill_data_copy$location.region[bill_data_copy$location.region == column_name] = "Europe";
}


# Zdruzimo Afriku
for (column_name in c("Lebanon","Egypt","Morocco","Algeria")){
  bill_data_copy$location.region[bill_data_copy$location.citizenship == column_name] = "Africa";
}


for (column_name in c("Sub-Saharan Africa")){
  bill_data_copy$location.region[bill_data_copy$location.region == column_name] = "Africa";
}


# zdruzimo Sjevernu Ameriku
for (column_name in c("North America")){
  bill_data_copy$location.region[bill_data_copy$location.region == column_name] = "North America";
}


# Zdruzimo Južnu Ameriku
for (column_name in c("Latin America")){
  bill_data_copy$location.region[bill_data_copy$location.region == column_name] = "South America";
}


# Zdruzimo Aziju
for (column_name in c("East Asia","South Asia")){
  bill_data_copy$location.region[bill_data_copy$location.region == column_name] = "Asia";
}
for (column_name in c("Saudi Arabia","Kuwait","United Arab Emirates","Israel","Turkey","Oman","Bahrain")
  bill_data_copy$location.region[bill_data_copy$location.citizenship == column_name] = "Asia";
}



bill_data_copy
```

```
tbl = table(bill_data_copy$location.region)
print(tbl)
```

```
##
##                 0         Africa           Asia         Europe North America
##                 1             43            699            697            992
## South America
##               182
```

##continent_frequency=transform(bill_data_copy,continent_frequency=ave(seq(nrow(bill_data_copy)),location.region

,FUN=length) df1=transform(bill_data_copy,continent_frequency=ave(seq(nrow(bill_data_copy)),location.region
,FUN=length)) df1

```r
df <- data.frame(continent=c("Europe", "Asia", "Africa","North America","South America"),
                 continent_frequency=c(697, 699, 43, 992, 182))
head(df)
```
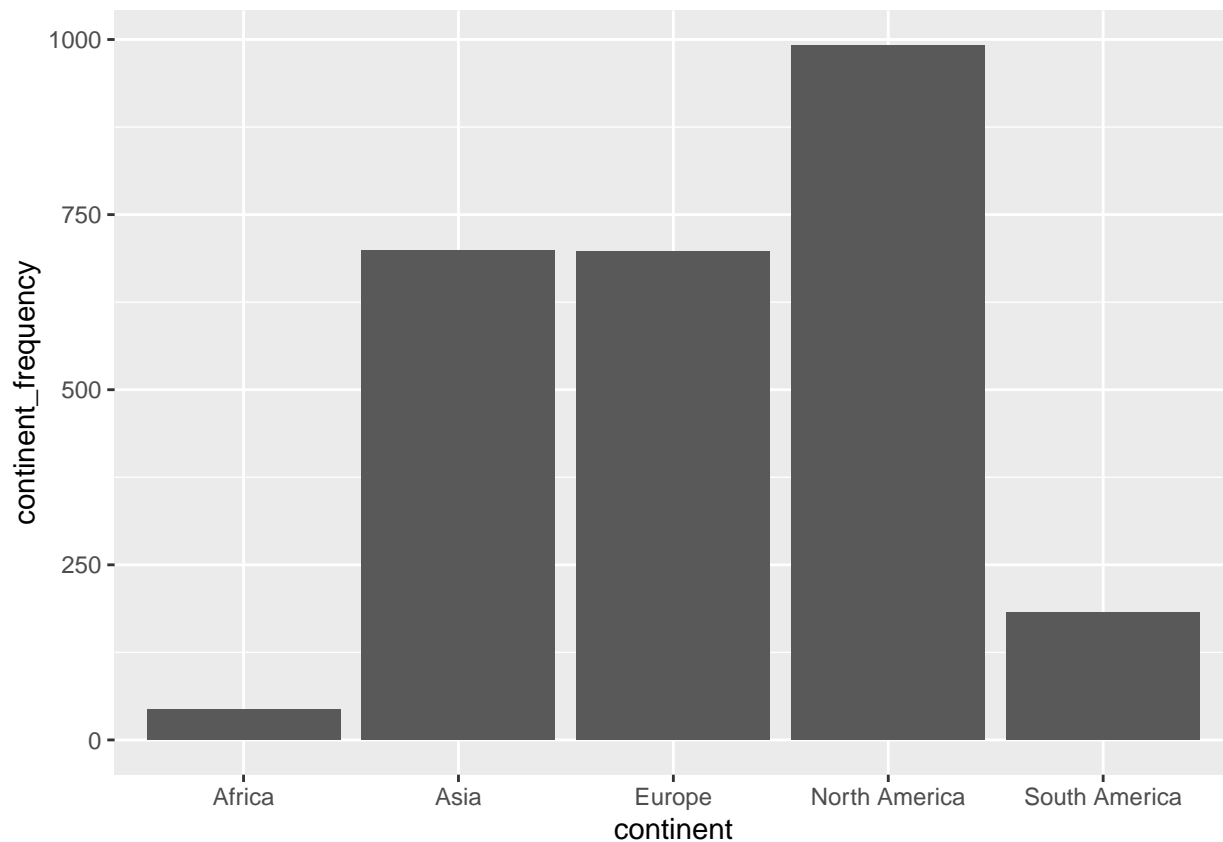
```
##         continent continent_frequency
## 1          Europe                 697
## 2            Asia                 699
## 3          Africa                  43
## 4   North America                 992
## 5   South America                 182
```

```r
library(ggplot2)


# Barplot
p<-ggplot(data=df, aes(x=continent, y=continent_frequency)) +
  geom_bar(stat="identity")
p
```



```r
box_edu <- ggplot(bill_data_copy %>% filter(!location.region=="0"), aes(x=location.region, y= wealth.wo
    geom_boxplot(alpha=0.7, ) + scale_y_log10() +
    stat_summary(fun=mean, geom="point", shape=20, size=7, color="black", fill="black") +
    theme(legend.position="none") + labs(x="Continent",y="Wealth")+
    scale_fill_brewer(name="Continent",palette="GnBu")
box_edu
```
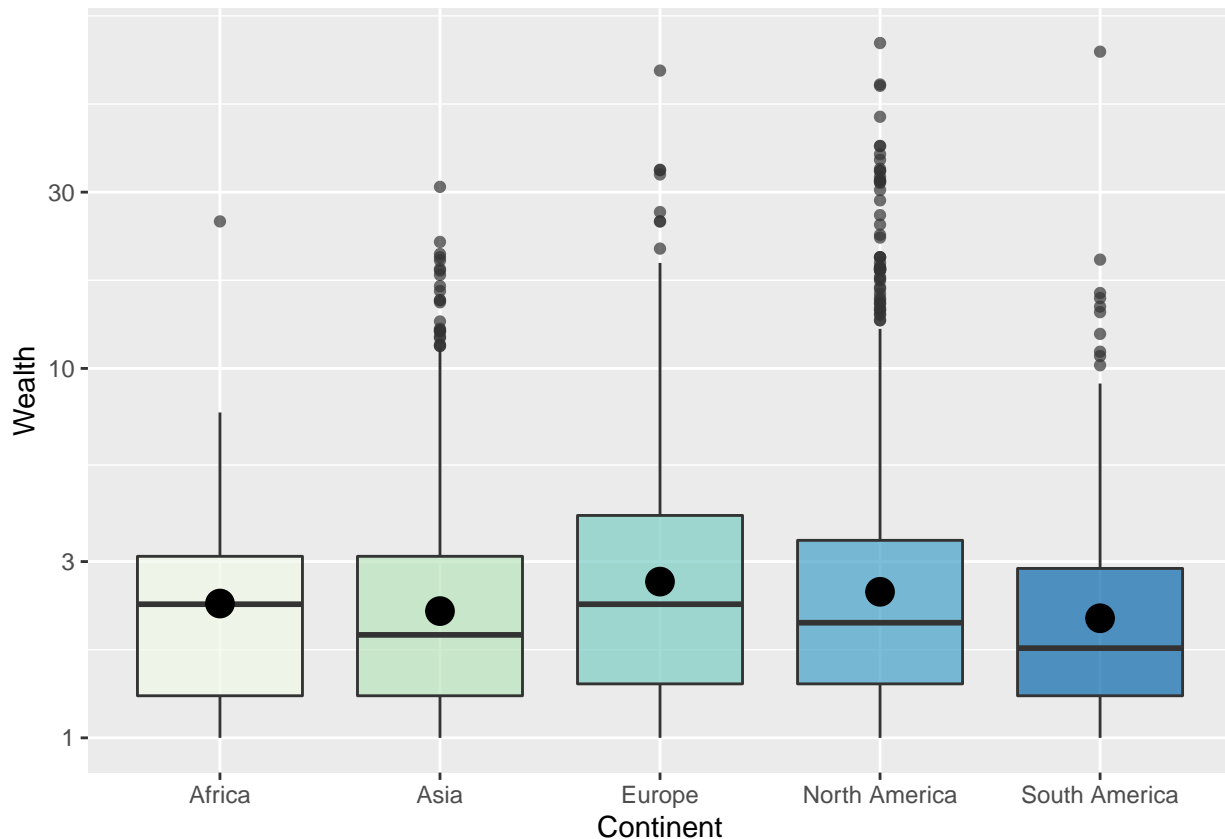
2. Jesu li milijarderi koji su nasljedili bogastvo statistički značajno bogatiji od onih koji nisu?

```
# Učitavanje podataka iz excel datoteke

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

```
## tracemem[0x6000030333a0 -> 0x6000030071e0]: lapply tbl_subset_row [.tbl_df [ eval eval withVisible w
```

```
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",]
```

```
## tracemem[0x6000030333a0 -> 0x60000303eae0]: lapply tbl_subset_row [.tbl_df [ eval eval withVisible w
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>, ...
```
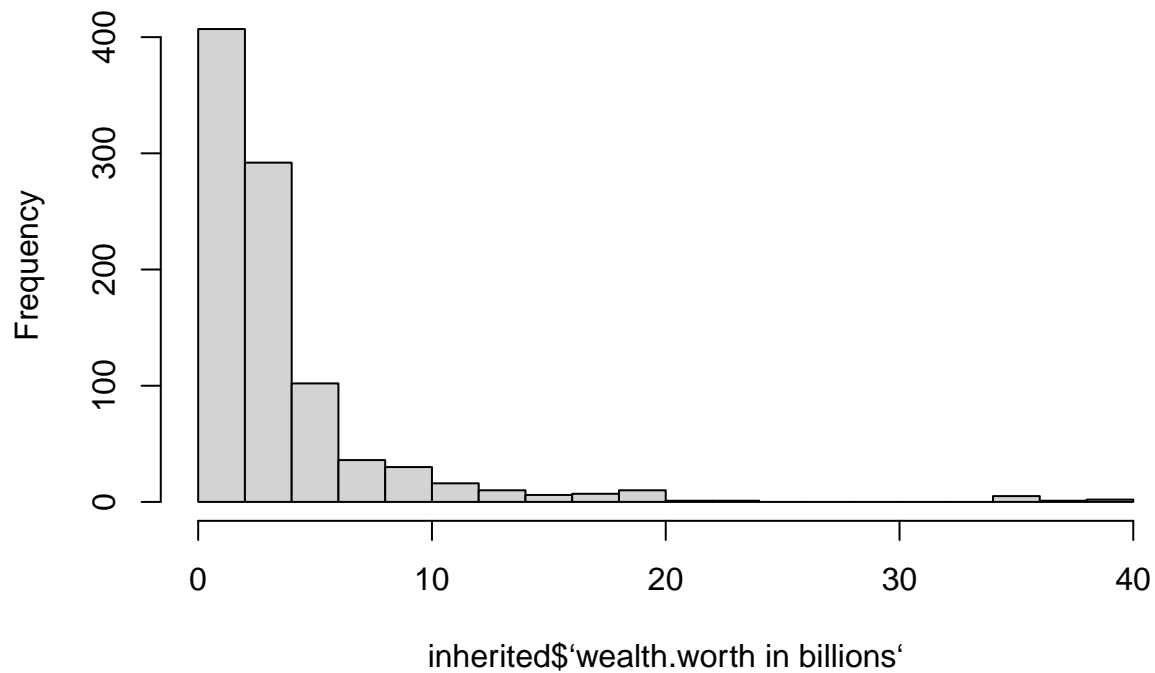
```
inherited_mean = mean(inherited$`wealth.worth in billions`)
print(inherited_mean)
```

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

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

Histogram of inherited$'wealth.worth in billions'

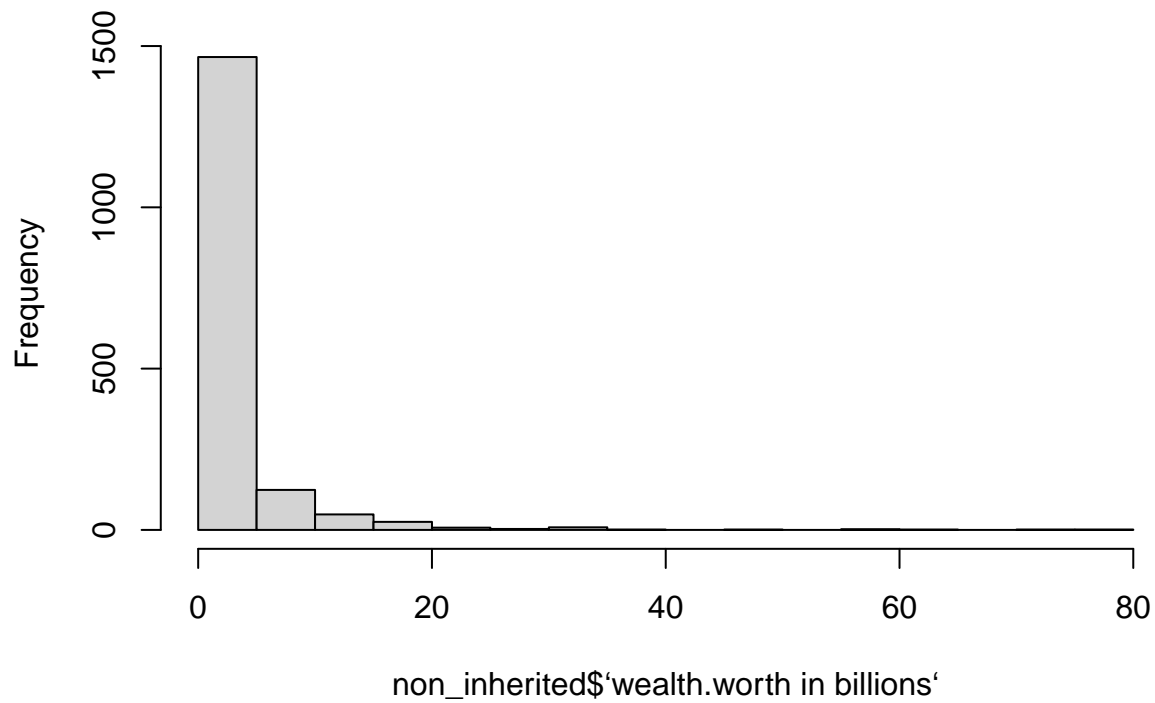inherited$'wealth.worth in billions'

```
non_inherited_mean = mean(non_inherited$`wealth.worth in billions`)
print(non_inherited_mean)
```
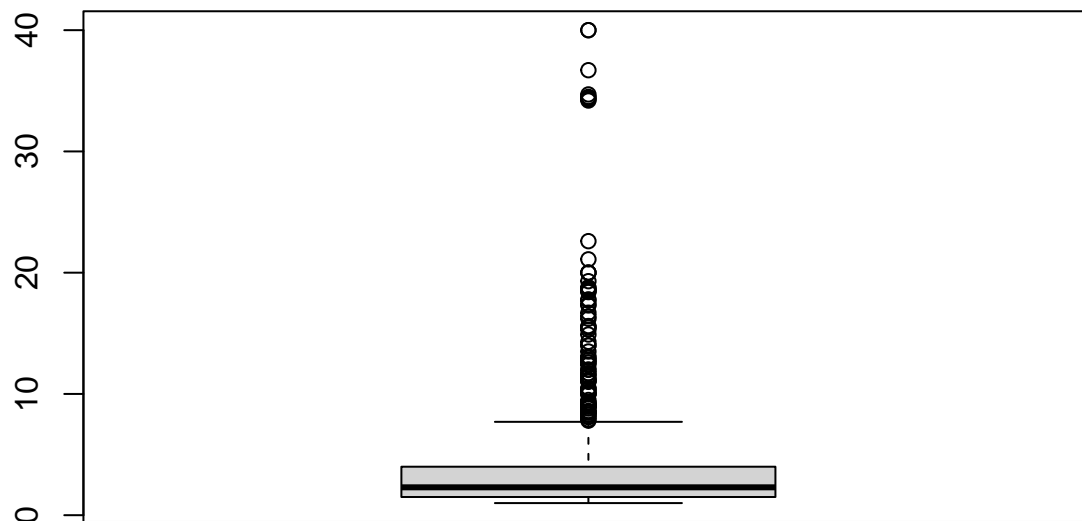
```
## [1] 3.411908
```

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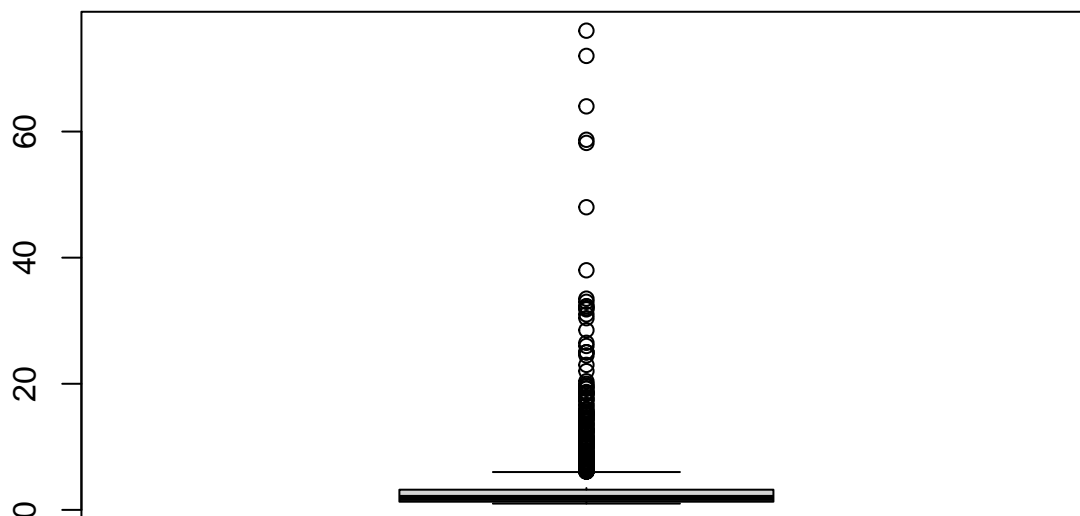
Histogram of non_inherited$'wealth.worth in billions'

non_inherited$'wealth.worth in billions'

```
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?

- je li dobro tu koristiti sve milijardere s popisa 2014 + milijarderi s prethodnih popisa (ako nisu na popisu iz 2014. godine)

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 naslijedili svoje bogatstvo. Također, zanimaju nas samo najnoviji milijarderi odnosno oni s popisa iz 2014. godine.

- kako prikazati trend kroz godine na grafu (dijagram paralelnih koordinata?)
- možda gledati razliku iz popisa 2014 i 2001, odnosno nove milijardere - pa napraviti raspodjelu industrija novonastalih milijardera

```
#
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)