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?

Deskriptivna analiza

Potrebno je učitati podatke.

```
# 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
```

```
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
##
##
##
```

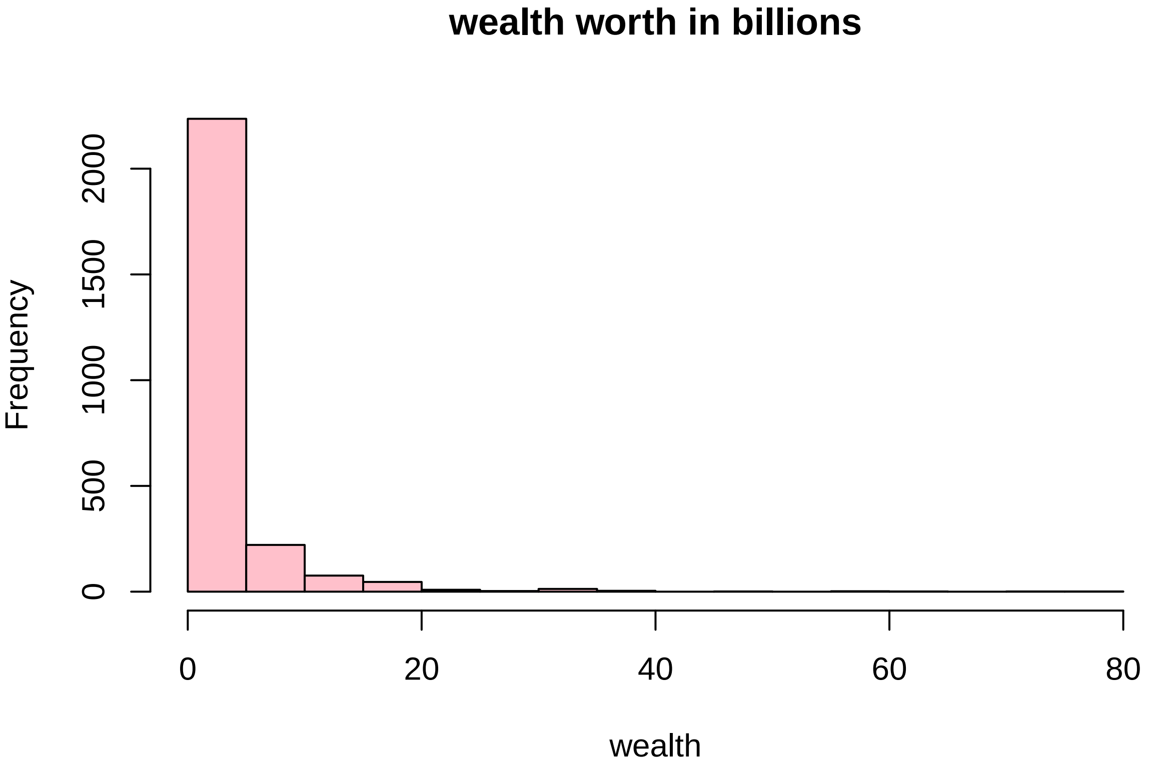
```
sapply(bill_data, class)
```

```
##                     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"
```

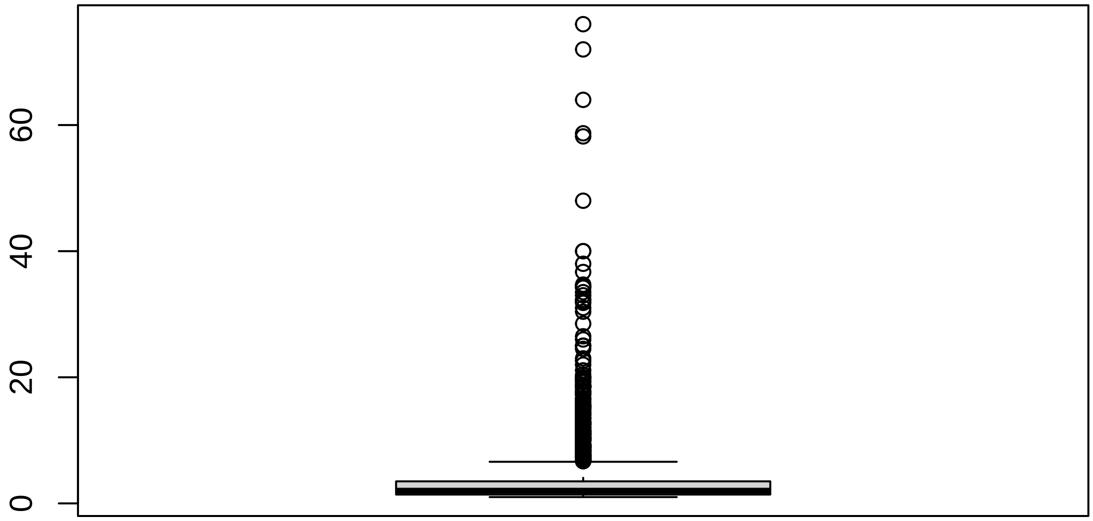
Naš dataset sastoji se od character i numeric varijabli.

Prvo promotrimo numeričke varijable.

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



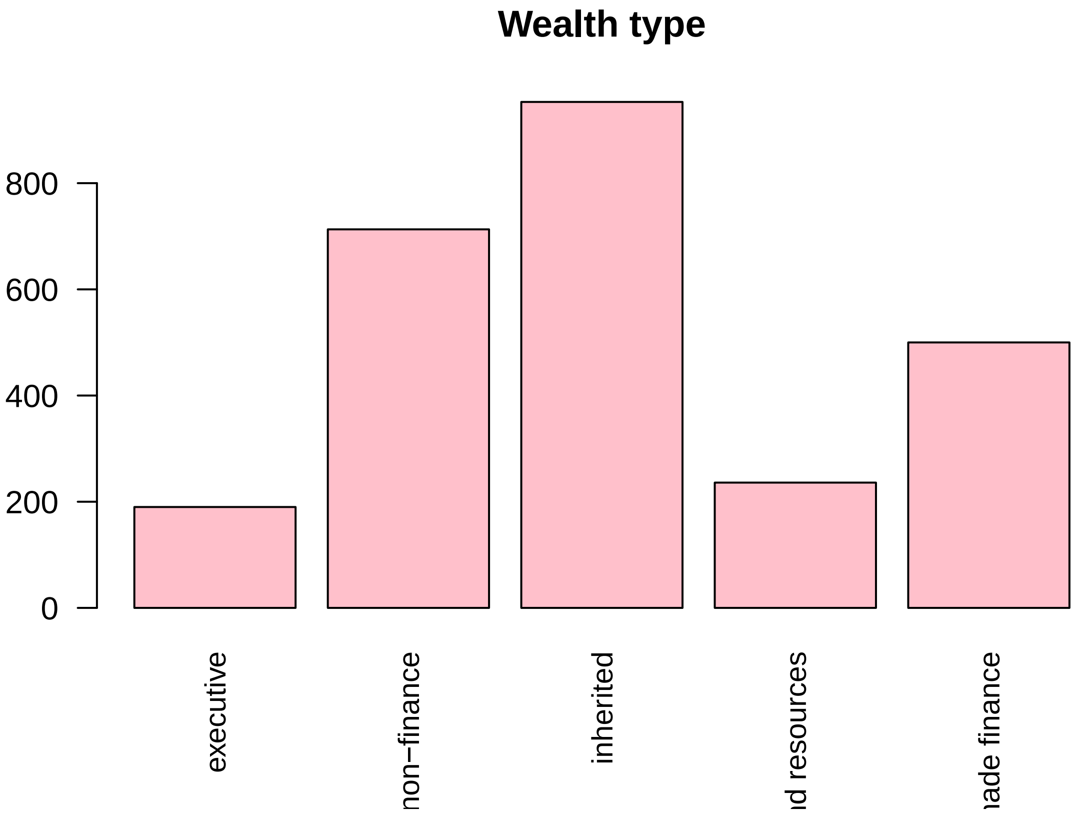
```
boxplot(bill_data$`wealth.worth in billions`)
```



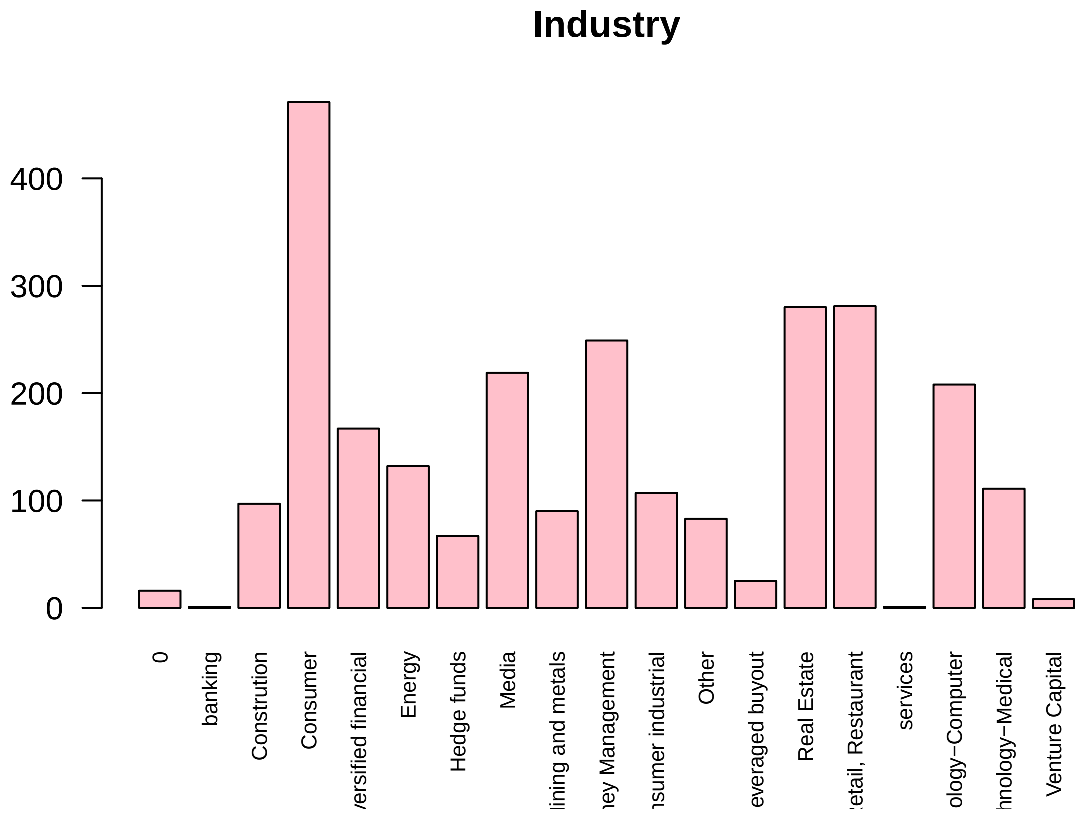
```
summary(bill_data$`wealth.worth in billions`)
```

```
##    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")
```



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



```
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 miljardi?

```
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
```

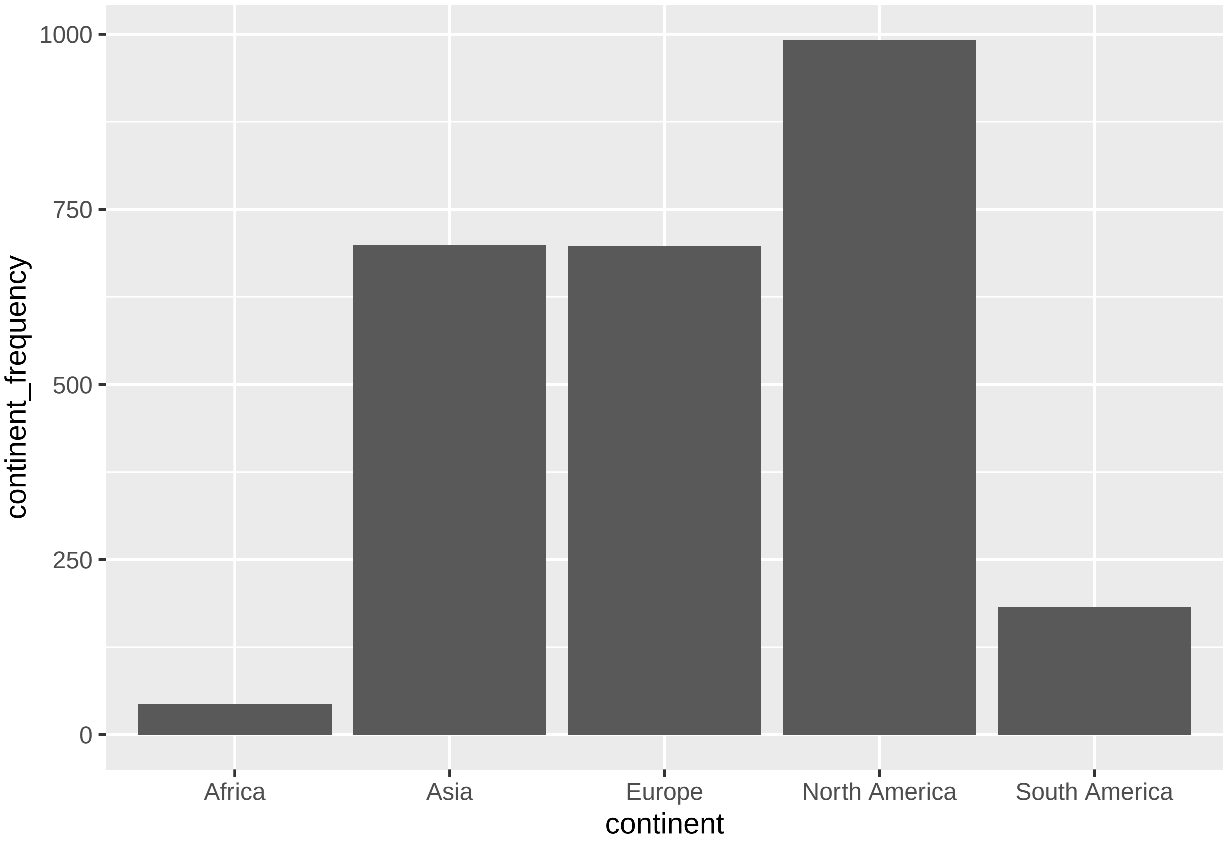
##continent_frequency=transform(bill_data_copy,continent_frequency=ave(seq(nrow(bill_data_copy)),location.region

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

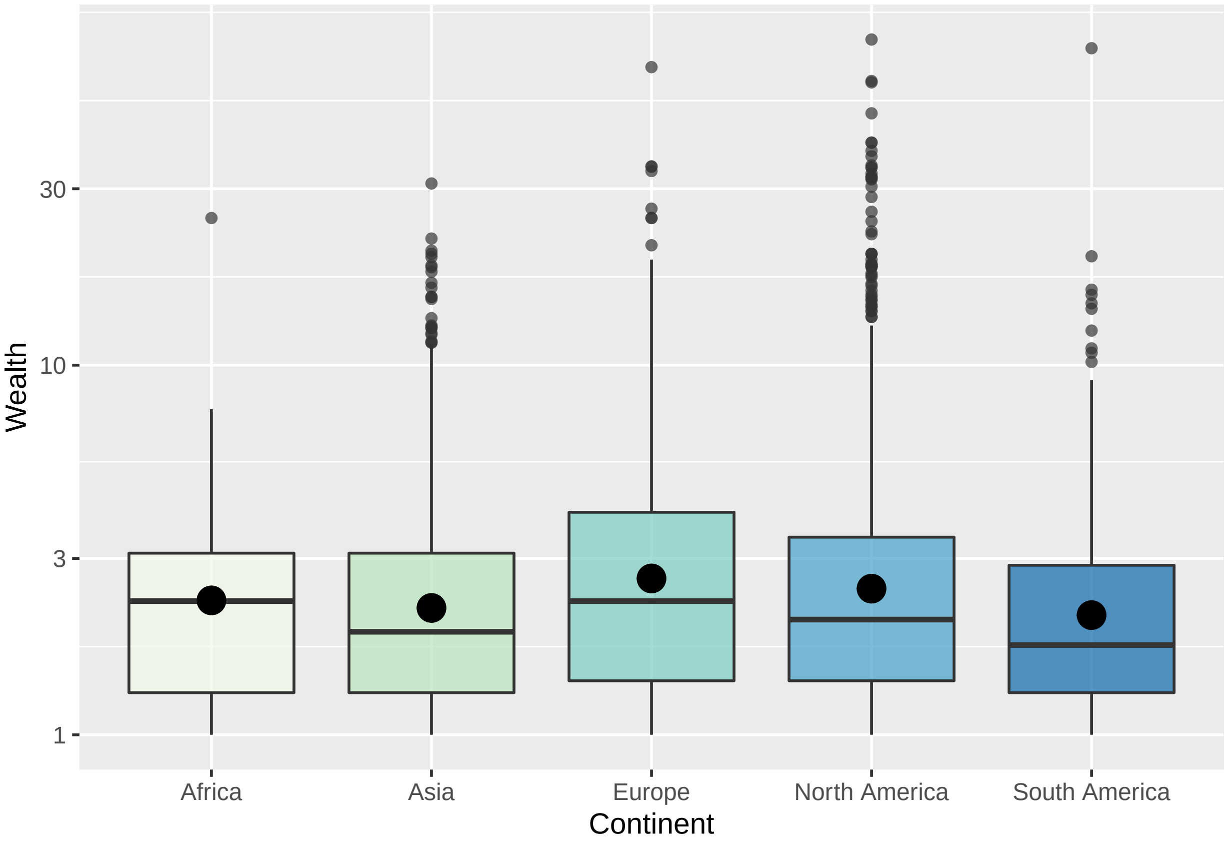
```
df <- data.frame(continent=c("Europe", "Asia", "Africa","North America","South America"),
                continent_frequency=c(697, 699, 43, 992, 182))
head(df)
```

```
##       continent continent_frequency
## 1        Europe                 697
## 2          Asia                 699
## 3        Africa                  43
## 4 North America                 992
## 5 South America                 182
```

```
library(ggplot2)


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



```
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
```



Pretpostavke ANOVA-e su:

- nezavisnost pojedinih podataka u uzorcima,
- normalna razdioba podataka,
- homogenost varijanci među populacijama.

Kad su veličine grupa podjednake, ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti i homogenosti varijanci. Ipak, dobro je provjeriti koliko su ta odstupanja velika.

Provjera normalnosti može se za svaku pojedinu grupu napraviti KS testom ili Lillieforsovom inačicom KS testa. U ovom slučaju razmatrat ćemo location.region kao varijablu koja određuje grupe (populacije) i wealth kao zavisnu varijablu.

```
# TODO: zakomentiraj ovu liniju ako ne želimo logaritmirati cijenu

wealth <- log(bill_data_copy$wealth.worth.in.billions, 2)

require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(wealth)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth
## D = 0.11777, p-value < 2.2e-16
```

```
lillie.test(wealth[bill_data_copy$location.region=='Africa'])
```

```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth[bill_data_copy$location.region == "Africa"]
## D = 0.12187, p-value = 0.112
```

```
lillie.test(wealth[bill_data_copy$location.region=='Europe'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth[bill_data_copy$location.region == "Europe"]
## D = 0.099476, p-value < 2.2e-16
```

```
lillie.test(wealth[bill_data_copy$location.region=='South America'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth[bill_data_copy$location.region == "South America"]
## D = 0.14997, p-value = 9.745e-11
```

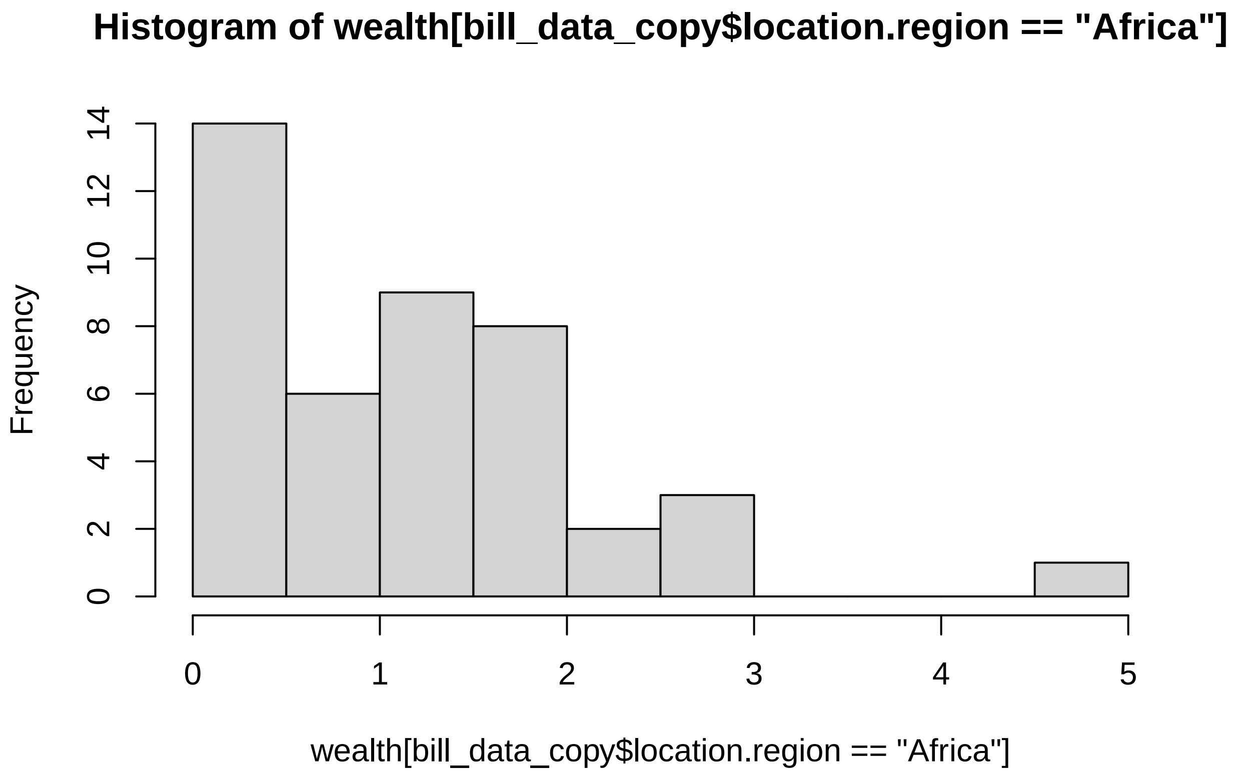
```
lillie.test(wealth[bill_data_copy$location.region=='North America'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth[bill_data_copy$location.region == "North America"]
## D = 0.12148, p-value < 2.2e-16
```

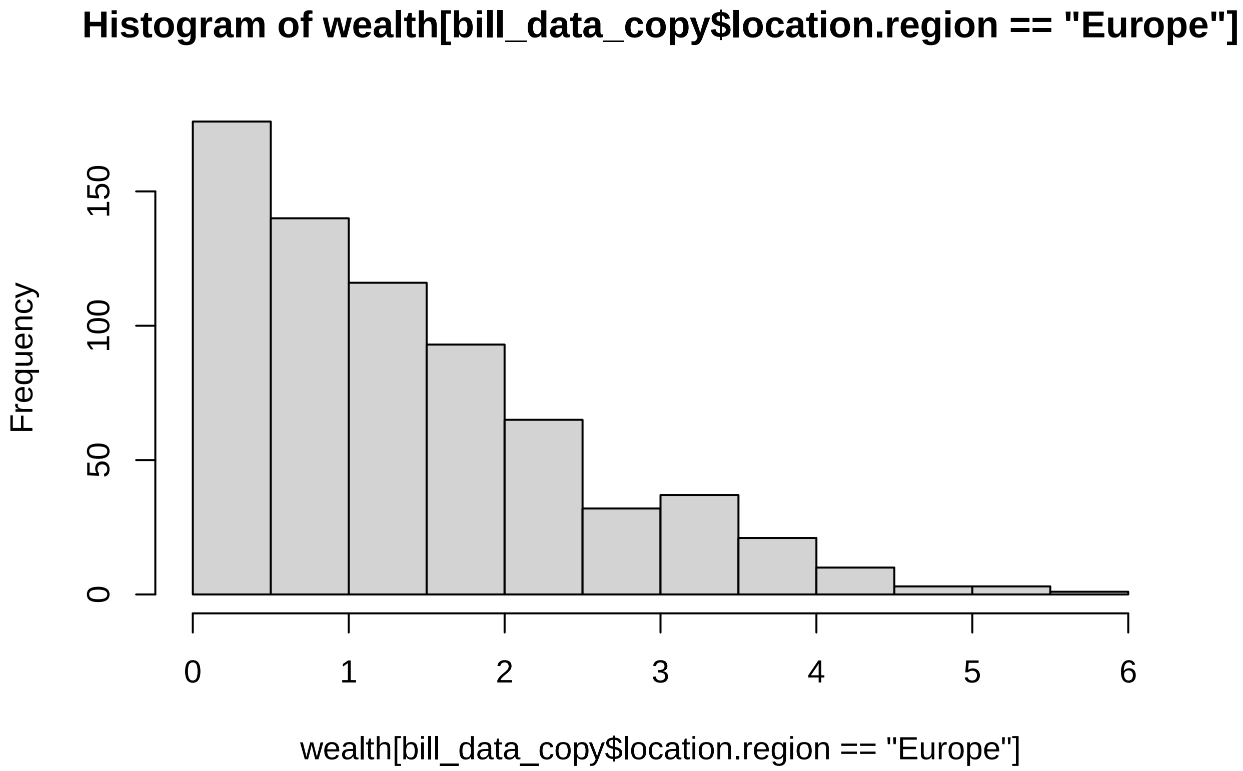
```
lillie.test(wealth[bill_data_copy$location.region=='Asia'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth[bill_data_copy$location.region == "Asia"]
## D = 0.12016, p-value < 2.2e-16
```

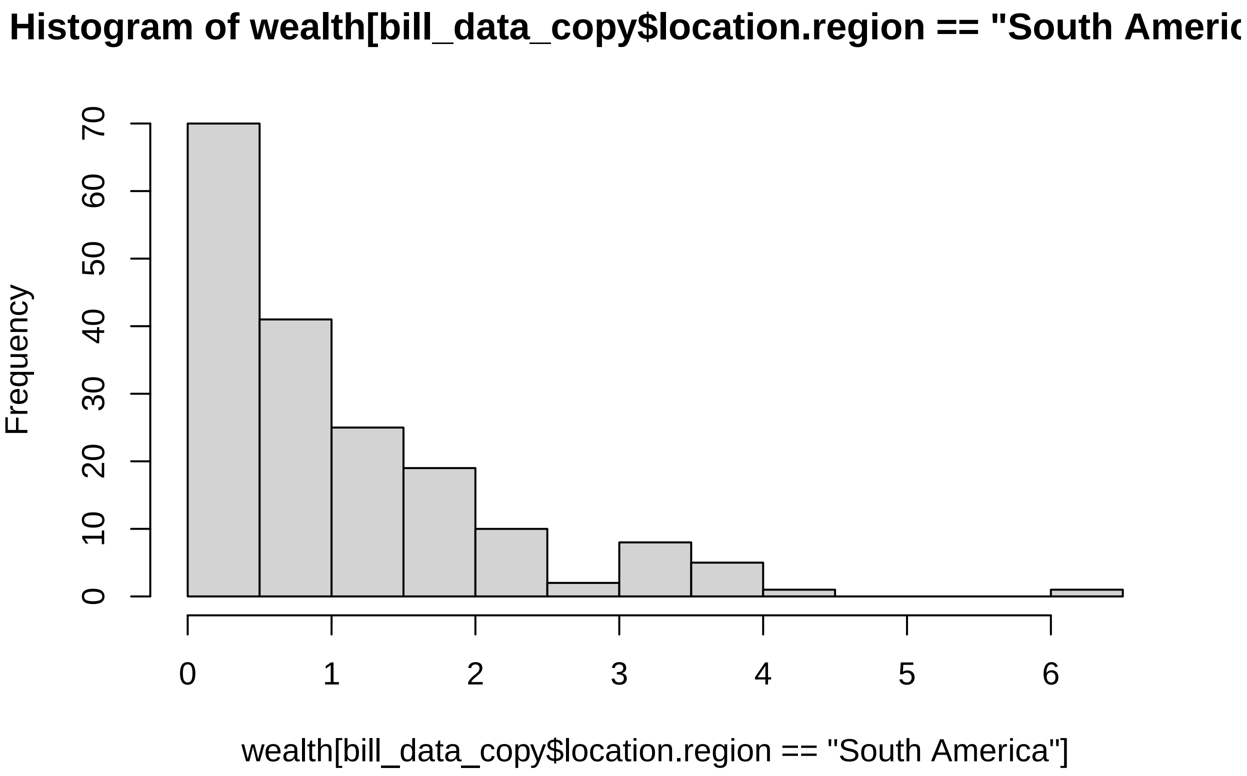
```
hist(wealth[bill_data_copy$location.region=='Africa'])
```



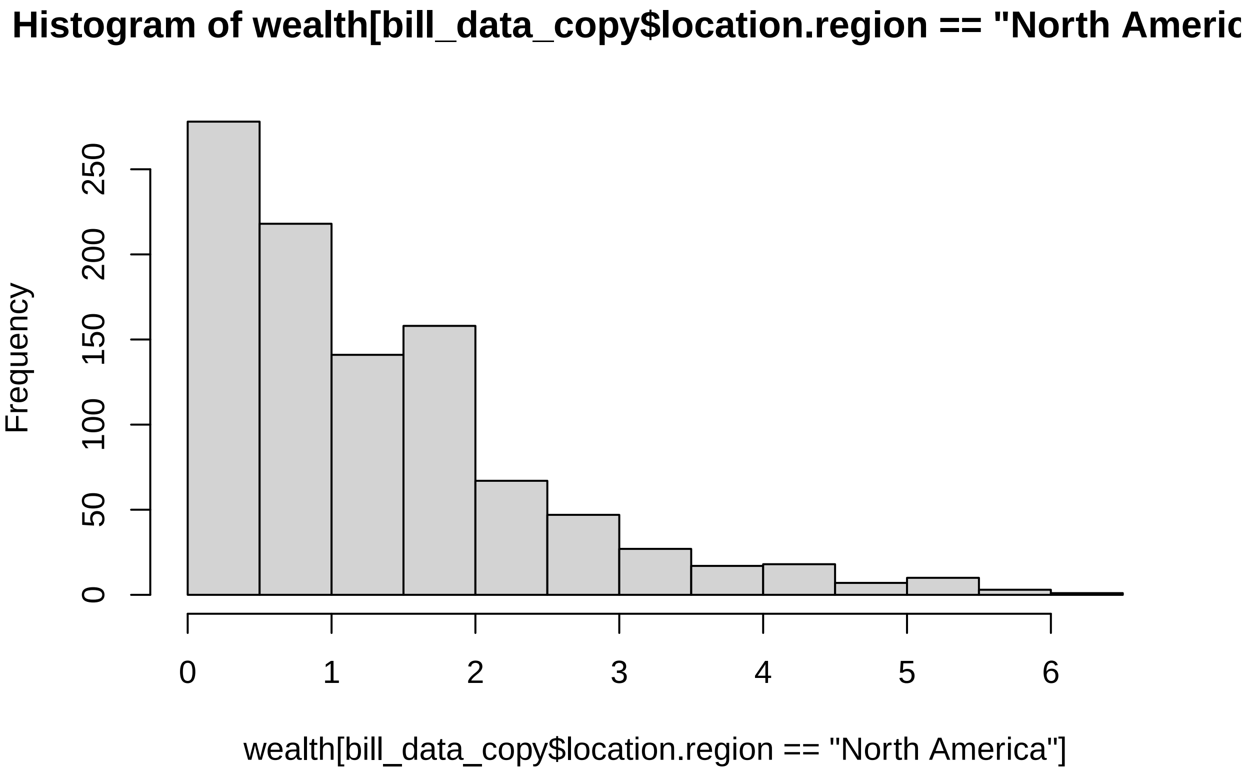
```
hist(wealth[bill_data_copy$location.region=='Europe'])
```



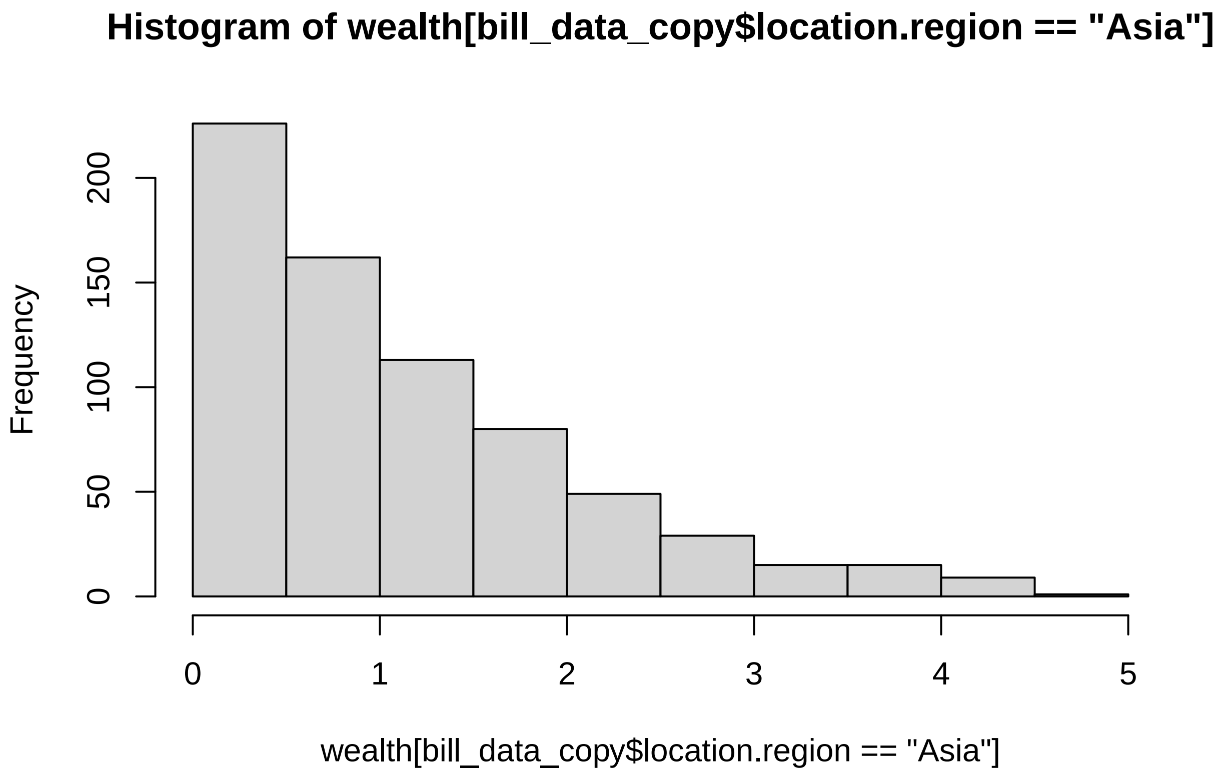
```
hist(wealth[bill_data_copy$location.region=='South America'])
```



```
hist(wealth[bill_data_copy$location.region=='North America'])
```



```
hist(wealth[bill_data_copy$location.region=='Asia'])
```



```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom

##bartlett.test(bill_data_copy$wealth.worth.in.billions ~ bill_data_copy$location.region)

var((wealth[bill_data_copy$location.region=='Africa']))
```

```
## [1] 0.8784496
```

```
var((wealth[bill_data_copy$location.region=='Asia']))
```

```
## [1] 0.9424432
```

```
var((wealth[bill_data_copy$location.region=='Europe']))
```

```
## [1] 1.196035
```

```
var((wealth[bill_data_copy$location.region=='North America']))
```

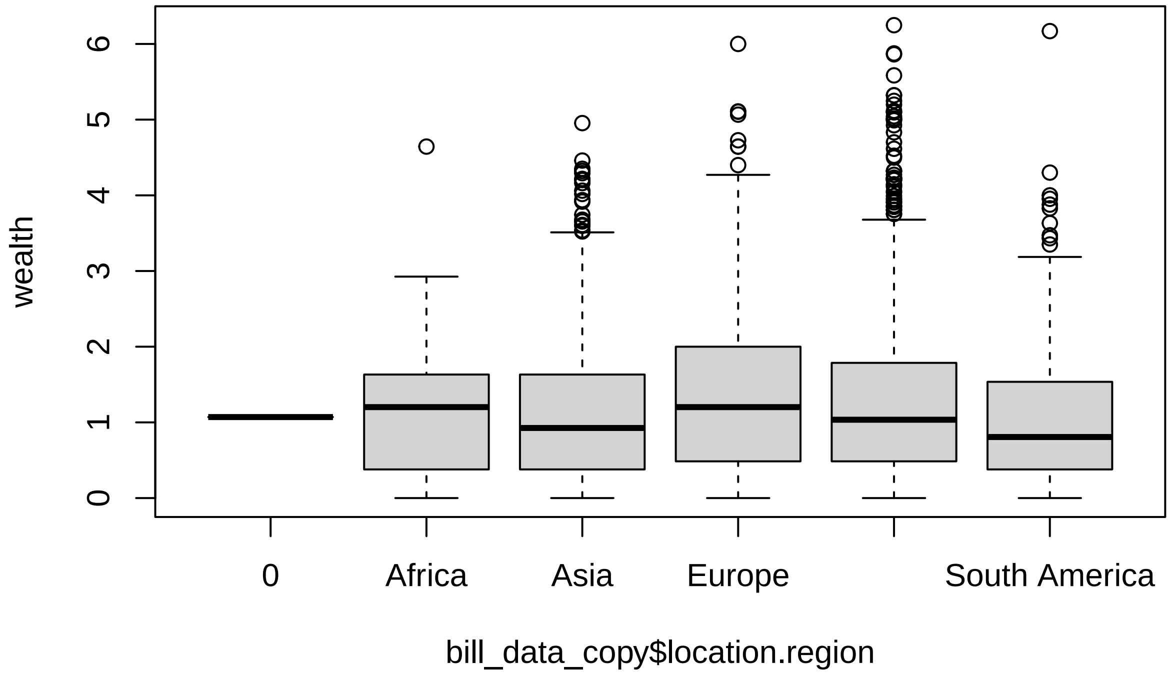
```
## [1] 1.265199
```

```
var((wealth[bill_data_copy$location.region=='South America']))
```

```
## [1] 1.076448
```

Provjerimo postoje li razlike u prihodima za različite razine školovanja klijenata.

```
# Graficki prikaz podataka
boxplot(wealth ~ bill_data_copy$location.region)
```



```
# Test
a = aov(wealth ~ bill_data_copy$location.region)
summary(a)
```

```
##                                  Df Sum Sq Mean Sq F value   Pr(>F)    
## bill_data_copy$location.region    5   33.4   6.689   5.862 2.15e-05 ***
## Residuals                      2608 2975.8   1.141                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Potrebno je pripremiti podatke za obradu, razdvojiti podatke iz tablice po polju how.inherited u dva slučaja: inherited (oni koju su nasljedili bogatstvo) i non_inherited (oni koji nisu nasljedili bogatstvo).

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

```
## tracemem[0x60000272e4c0 -> 0x6000027142a0]: lapply tbl_subset_row [.tbl_df [ eval eval withVisible wi
```

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

```
## tracemem[0x60000272e4c0 -> 0x600002715ea0]: lapply tbl_subset_row [.tbl_df [ eval eval withVisible wi
```

Zatim je potrebno izračunati srednju vrijednost (mean) posebno za svaki slučaj uzimajući u obzir polje worth.in billions.

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

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

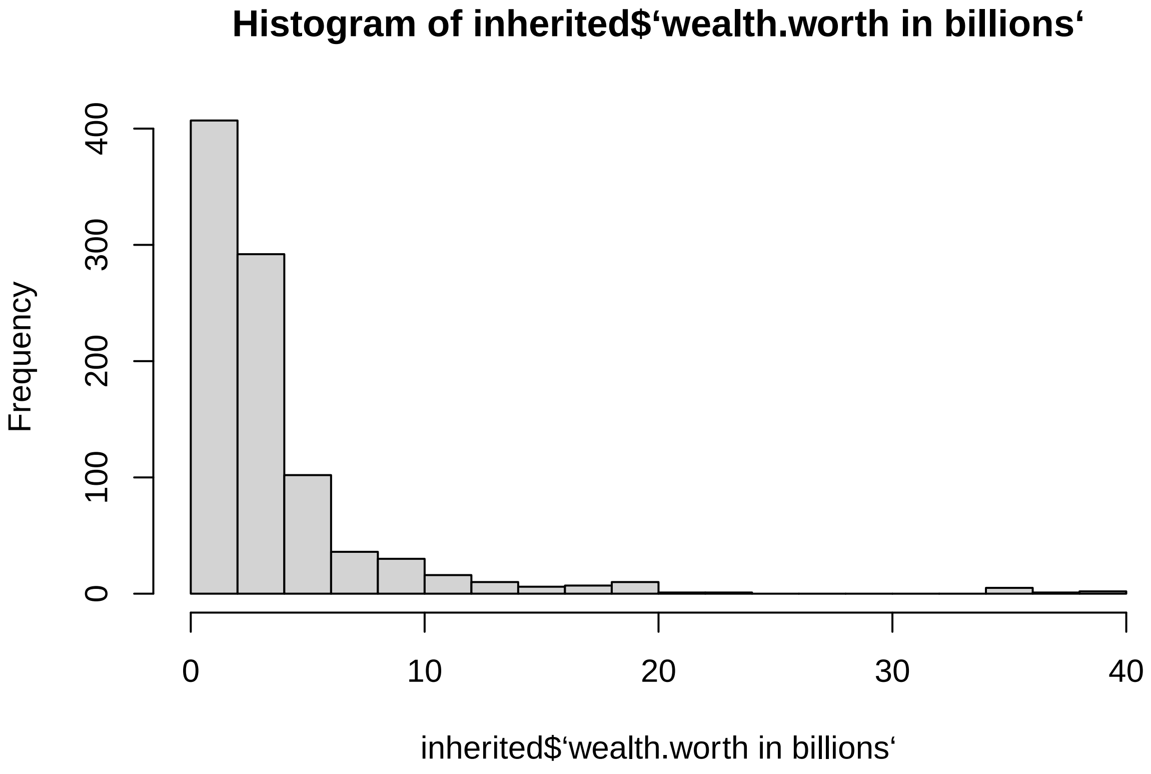
```
non_inherited_mean = mean(non_inherited$`wealth.worth in billions`)
print(non_inherited_mean)
```

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

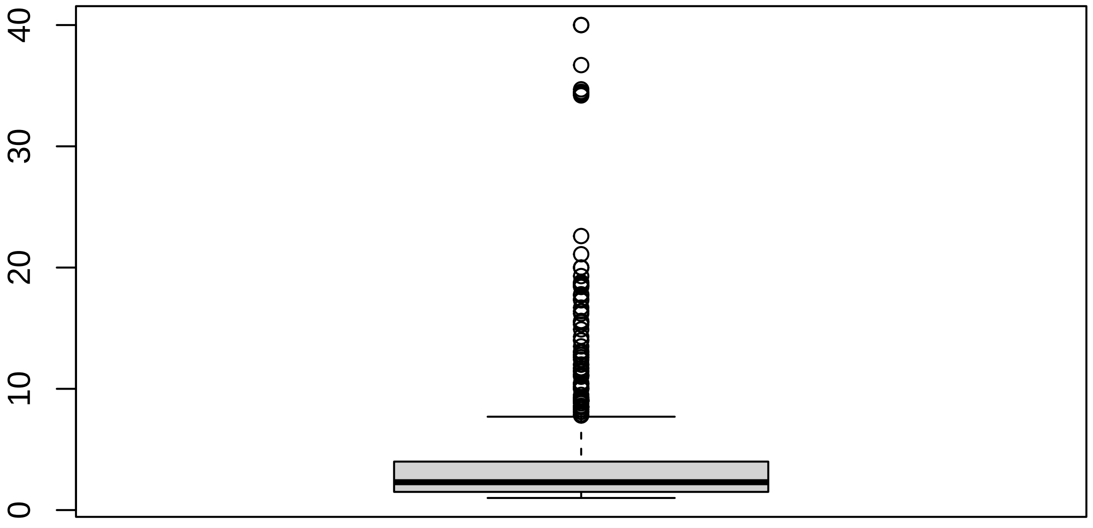
Na temelju male razlike u srednjim vrijednostima, ne postoje indikacije da su milijarderi koji su nasljedili bogatstvo statistički značajno bogatiji od onih koji nisu. No, navedeno je potrebno provjeriti.

Kako bi bolje vizualizirali podatke crtamo histogram i box plot za svaki od slučaja:

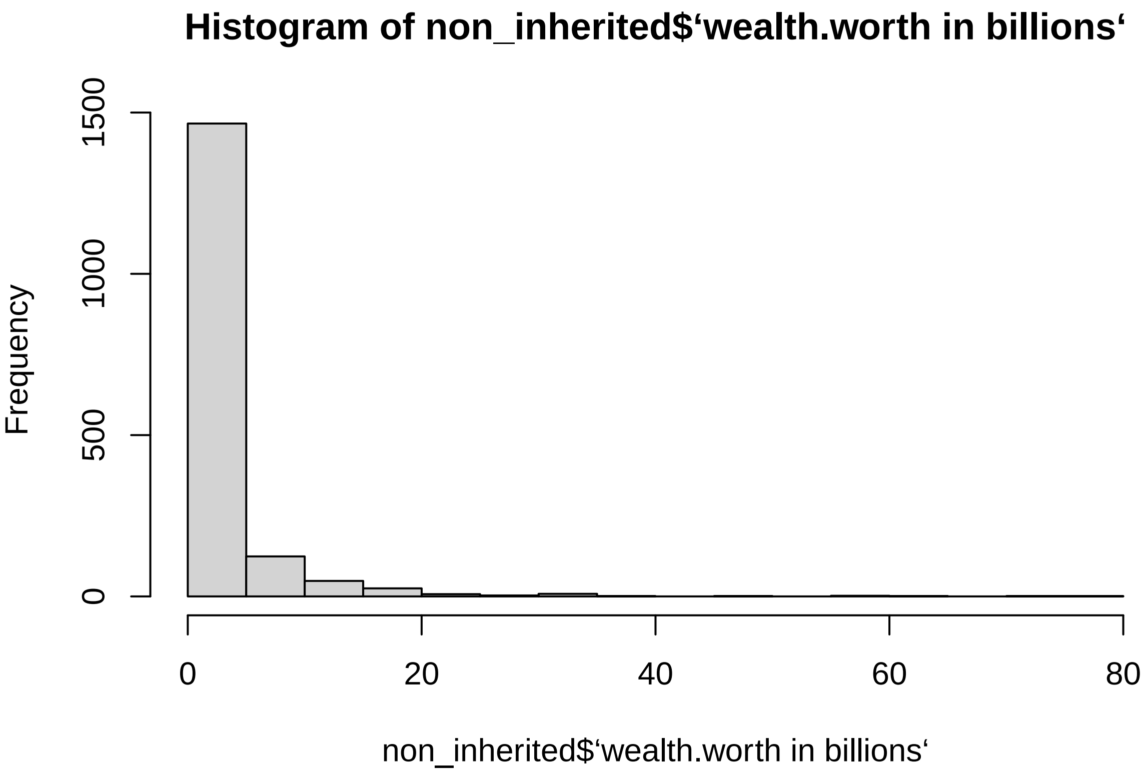
```r
hist(inherited$`wealth.worth in billions`, breaks = 20)
```



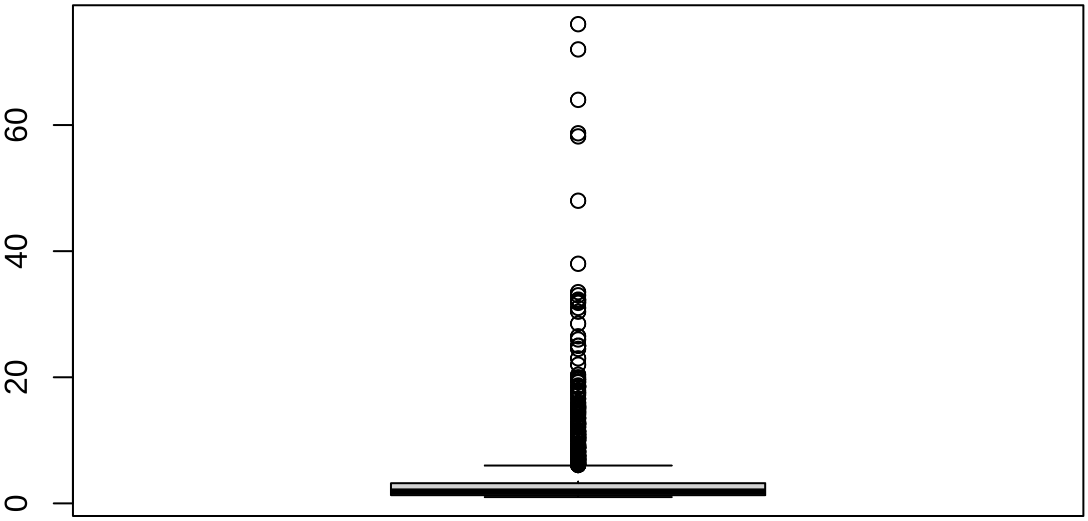
```r
boxplot(inherited$`wealth.worth in billions`)
```



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



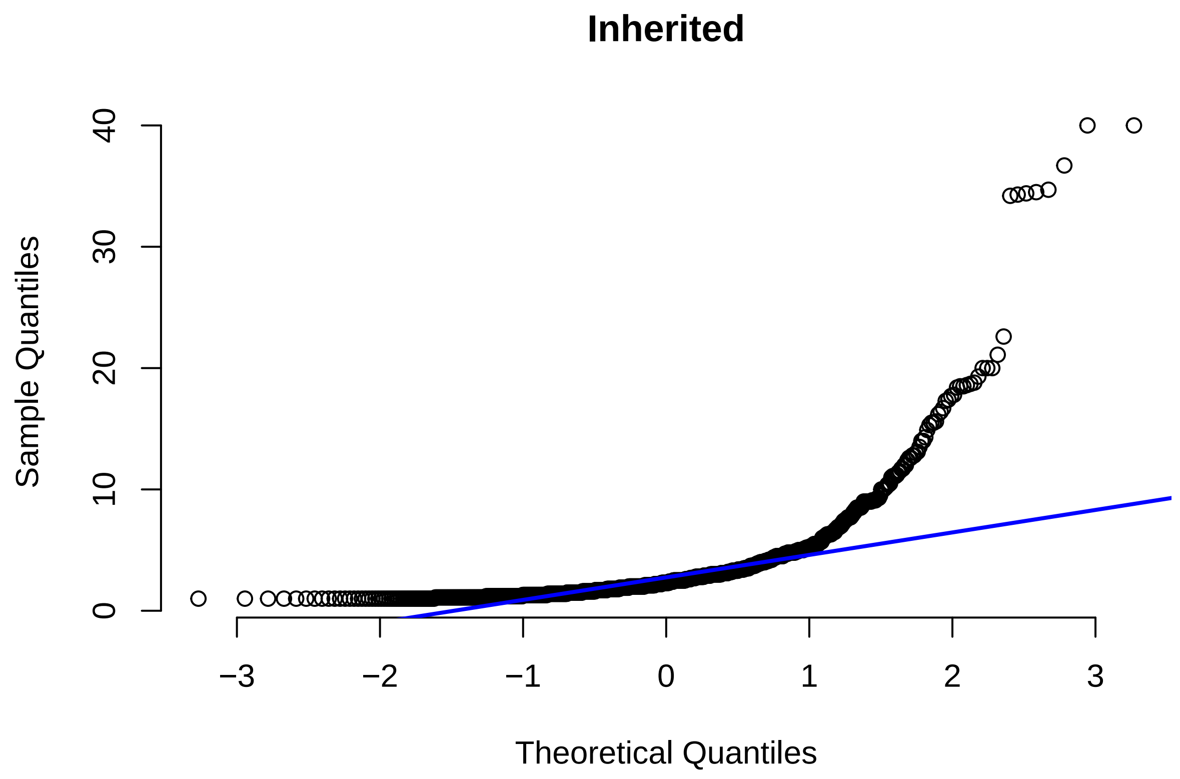
```
boxplot(non_inherited$`wealth.worth in billions`)
```



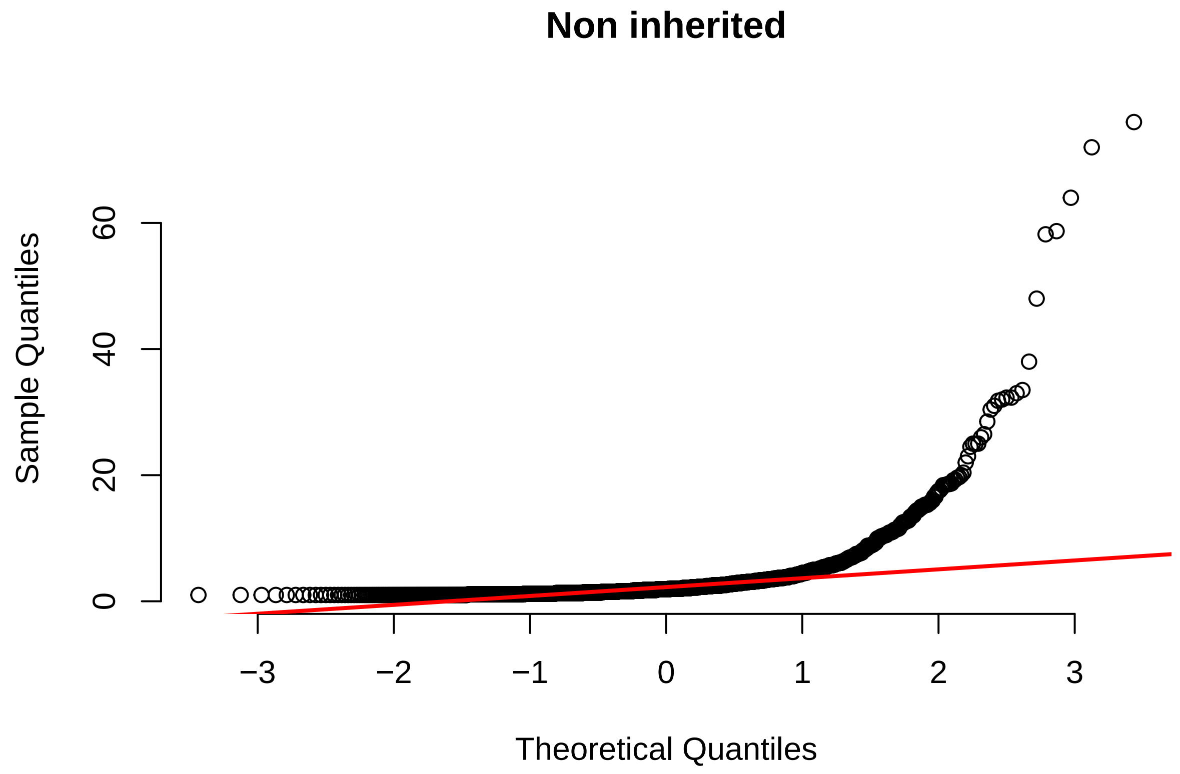
Iz prikazane vizualizacije uočavamo kako se podaci ne ravnaju po normalnoj distribuciji.

Što se može bolje vidjeti sa sljedećih prikaza:

```
qqnorm(inherited$`wealth.worth in billions`, pch = 1, frame = FALSE,main='Inherited')
qqline(inherited$`wealth.worth in billions`, col = "blue", lwd = 2)
```



```
qqnorm(non_inherited$`wealth.worth in billions`, pch = 1, frame = FALSE,main='Non inherited')
qqline(non_inherited$`wealth.worth in billions`, col = "red", lwd = 2)
```



Ipak, uočeno je potrebno dodatno ispitati koristeći Kolmogorov–Smirnov test kojim se utvrđuje ravna li se distribucija po normalnoj razdiobi.

```
ks.test(inherited$`wealth.worth in billions`, y="pnorm")
```

```
## Warning in ks.test(inherited$`wealth.worth in billions`, y = "pnorm"): ties
## should not be present for the Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  inherited$`wealth.worth in billions`
## D = 0.84134, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(non_inherited$`wealth.worth in billions`, y="pnorm")
```

```
## Warning in ks.test(non_inherited$`wealth.worth in billions`, y = "pnorm"): ties
## should not be present for the Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  non_inherited$`wealth.worth in billions`
## D = 0.84134, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Iz dobivenih p vrijednosti u oba slučaja odbacujemo mogućnost da se distribucije ravnaju po normalnoj razdiobi.

Time je potvrđena pretpostavka da se podaci ne ravnaju po normalnoj distribuciji.

Potrebno je koristiti neparametarski test Mann–Whitney U test, koji se koristi kada se podaci se ravnaju po istim distribucijama (obje distribucije su nakošene u desno) i uzorci su nezavisni iz jedne i druge populacije (jedna osoba ne može nasljediti i nenasljediti bogatstvo).

Hipoteze glase:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

```
wilcox.test(inherited_mean, non_inherited_mean, alt = "greater")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  inherited_mean and non_inherited_mean
## W = 1, p-value = 0.5
## alternative hypothesis: true location shift is greater than 0
```

Zbog p-vrijednost jednake 0.5, na temelju značajnosti od 50% ne možemo odbaciti H_0 hipotezu o jednakosti prosječnih vrijednosti bogatstva u korist H_1, odnosno možemo reći da milijarderi koji su nasljedili bogatstvo nisu statistički značajno bogatiji od onih koji nisu.

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)

```
# bill_data

# Izbacujemo stupce:
# name
```

```r
# company.name
# rank
# location.gdp, više od pola vrijednosti su 0 (netočan podatak)
# location.coutnry.code i location.citizenship a koristimo location.region koji je veće granulacije
# wealth.how.from emerging, wealth.how.was founder, wealth.how.was political su konstantne varijable
# company.sector jer ima previše različitih vrijednosti, koje kad bi one hot encodali bi dali previše s

exclude_cols = c("name", "company.name", "rank", "location.gdp", "location.country code", "location.citi

# exclude columns and sort
bill_data_clean <- bill_data %>% select(-one_of(exclude_cols)) %>% arrange(year)

# to lowercase for consistency
bill_data_clean[["company.relationship"]] <- tolower(bill_data_clean[["company.relationship"]] )

# remove invalid data
bill_data_clean <- bill_data_clean %>% filter(demographics.age > 0)
bill_data_clean <- bill_data_clean %>% filter(!location.region == "0")

# inflation rate $1.00 (1996) -> $1.51 (2014), +50.9%
# inflation rate $1.00 (2001) -> $1.34 (2014), +33.7%
bill_data_clean[bill_data_clean$year == "1996", "wealth.worth in billions"] <- bill_data_clean[bill_data
bill_data_clean[bill_data_clean$year == "2001", "wealth.worth in billions"] <- bill_data_clean[bill_data

# Iskoristili smo godinu da ažuriramo cijene (inflacija), sad ju odbacujemo
bill_data_clean <- bill_data_clean %>% select(., -year)

# merge similar roles to avoid 1 column = 1 row data
bill_data_clean$company.relationship <- gsub(".*\b(owner)\b.*", "owner", bill_data_clean$company.relatio
bill_data_clean$company.relationship <- gsub(".*(ceo|chief executive officeor|chief executive officer|ch
bill_data_clean$company.relationship <- gsub(".*(founder).*", "founder", bill_data_clean$company.relatio
bill_data_clean$company.relationship <- gsub(".*(chair|chari).*", "chairman", bill_data_clean$company.re
bill_data_clean$company.relationship <- gsub(".*(director).*", "director", bill_data_clean$company.relat
bill_data_clean$company.relationship <- gsub(".*(head).*", "head", bill_data_clean$company.relationship)
bill_data_clean$company.relationship <- gsub(".*(president).*", "president", bill_data_clean$company.rel

# drop small amount of rows with na values
bill_data_clean <- bill_data_clean %>% drop_na()

# split dataset to numeric and categorical (non-ordinal)
bill_categorical <- bill_data_clean %>% select(where(is_character))
bill_numeric <- bill_data_clean %>% select(where(is.numeric))

# one hot encode categorical data
bill_categorical_onehot = dummy_cols(bill_categorical, remove_first_dummy = TRUE, remove_selected_column

# filter indicators with 5 or more rows, indicators with less than 5 would cause problems
bill_categorical_onehot <- bill_categorical_onehot[, colSums(bill_categorical_onehot) > 5]

# concat numerical and categorical columns
bill_data_clean <- bind_cols(bill_numeric, bill_categorical_onehot)
```

```r
# remove variables which strongly and linearly correlate
correlation_threshold = 0.9
```

```
tmp <- cor(bill_data_clean)
tmp[upper.tri(tmp)] <- 0
diag(tmp) <- 0  # clean diagonal which is always 1
bill_data_clean <- bill_data_clean[, apply(tmp,2,function(x) all(x<= correlation_threshold))]

# remove outliers
# TODO: zakomentiraj ovu liniju ako ne želimo removeati outliere
bill_data_clean <- remove_outliers(bill_data_clean, bill_data_clean$`wealth.worth in billions`)

# extract y column for later use
wealth <- bill_data_clean$`wealth.worth in billions`

# TODO: zakomentiraj ovu liniju ako ne želimo logaritmirati cijenu
# wealth <- log(wealth, 2)
```

```
# x setup, y = wealth
normalized<-function(y) {
  x<-y[!is.na(y)]
  x<-(x - min(x)) / (max(x) - min(x))
  y[!is.na(y)]<-x
  return(y)
}

# `wealth.how.industry_Retail, Restaurant` casues fitting issues
exclude_cols = c("wealth.worth in billions", "wealth.how.industry_Retail, Restaurant")
x <- bill_data_clean %>% select(-one_of(exclude_cols))
x[, c("company.founded", "demographics.age")] <- apply(x[, c("company.founded", "demographics.age")] , 2
x <- x[,order(colnames(x))]
```

```
model_all_vars <- lm(wealth ~ . , x)
summary(model_all_vars)
```

```
##
## Call:
## lm(formula = wealth ~ ., data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5317 -1.0244 -0.3629  0.6162  5.0657
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                             1.61419    1.17790
## company.founded                                         0.41979    0.60953
## company.relationship_chairman                          -0.28734    0.20051
## company.relationship_director                          -0.80622    0.68723
## company.relationship_founder                           -0.34965    0.17998
## company.relationship_investor                          -0.51311    0.29097
## company.relationship_owner                             -0.75785    0.22154
## company.relationship_president                          0.12327    0.42778
## company.type_aquired                                   -1.29221    0.41931
## company.type_new                                       -1.37453    0.40406
## company.type_privatization                             -1.61248    0.48589
## company.type_subsidiary                                -2.58826    0.73477
## demographics.age                                       1.26004    0.23747
## demographics.gender_male                               0.19710    0.13437
## location.region_Europe                                 0.16762    0.10820
## `location.region_Latin America`                       -0.46210    0.15787
## `location.region_Middle East/North Africa`            -0.20107    0.18424
## `location.region_North America`                        0.12757    0.09931
## `location.region_South Asia`                          -0.27365    0.21188
## `location.region_Sub-Saharan Africa`                   0.60954    0.36632
## wealth.how.category_Financial                          0.57625    0.60382
## `wealth.how.category_New Sectors`                     -2.28107    1.64451
## `wealth.how.category_Non-Traded Sectors`               0.93174    0.49731
## `wealth.how.category_Resource Related`                -0.95728    0.83705
## `wealth.how.category_Traded Sectors`                  -0.71926    0.69180
## wealth.how.industry_Constrution                       -0.39673    0.20547
## wealth.how.industry_Consumer                           1.58716    0.84381
## `wealth.how.industry_Diversified financial`            0.46262    0.68373
## wealth.how.industry_Energy                             1.54698    0.96578
## `wealth.how.industry_Hedge funds`                      0.35839    0.69779
## wealth.how.industry_Media                              0.31209    0.15984
## `wealth.how.industry_Mining and metals`                1.59869    0.97105
## `wealth.how.industry_Money Management`                 0.55354    0.67792
## `wealth.how.industry_Non-consumer industrial`          1.49457    0.86078
## wealth.how.industry_Other                              0.50130    0.51415
## `wealth.how.industry_Private equity/leveraged buyout`  1.12879    0.73749
## `wealth.how.industry_Real Estate`                      0.41929    0.67375
## `wealth.how.industry_Technology-Computer`              3.34813    1.71775
## `wealth.how.industry_Technology-Medical`               2.79229    1.70488
## `wealth.how.industry_Venture Capital`                 -0.11704    0.85116
## `wealth.how.inherited_4th generation`                  0.22361    0.23883
## `wealth.how.inherited_5th generation or longer`       -0.15492    0.35896
## wealth.how.inherited_father                            0.19327    0.14990
## `wealth.how.inherited_not inherited`                  -0.11821    0.75791
## `wealth.how.inherited_spouse/widow`                   -0.28484    0.29805
## `wealth.type_founder non-finance`                      0.56035    0.19385
## wealth.type_inherited                                  0.41711    0.76615
## `wealth.type_privatized and resources`                 0.59664    0.23540
## `wealth.type_self-made finance`                        0.11468    0.21904
##                                                       t value Pr(>|t|)
## (Intercept)                                              1.370 0.170722
## company.founded                                          0.689 0.491087
## company.relationship_chairman                           -1.433 0.152012
## company.relationship_director                           -1.173 0.240881
## company.relationship_founder                            -1.943 0.052190 .
## company.relationship_investor                           -1.763 0.077981 .
## company.relationship_owner                              -3.421 0.000637 ***
## company.relationship_president                           0.288 0.773248
## company.type_aquired                                    -3.082 0.002087 **
## company.type_new                                        -3.402 0.000683 ***
## company.type_privatization                              -3.319 0.000921 ***
## company.type_subsidiary                                 -3.523 0.000437 ***
## demographics.age                                         5.306 1.25e-07 ***
## demographics.gender_male                                 1.467 0.142586
## location.region_Europe                                   1.549 0.121499
## `location.region_Latin America`                         -2.927 0.003461 **
## `location.region_Middle East/North Africa`             -1.091 0.275264
## `location.region_North America`                         1.285 0.199097
## `location.region_South Asia`                           -1.292 0.196674
## `location.region_Sub-Saharan Africa`                    1.664 0.096288 .
## wealth.how.category_Financial                           0.954 0.340031
## `wealth.how.category_New Sectors`                      -1.387 0.165576
## `wealth.how.category_Non-Traded Sectors`                1.874 0.061140 .
## `wealth.how.category_Resource Related`                 -1.144 0.252916
## `wealth.how.category_Traded Sectors`                   -1.040 0.298608
## wealth.how.industry_Constrution                        -1.931 0.053649 .
## wealth.how.industry_Consumer                            1.881 0.060130 .
## `wealth.how.industry_Diversified financial`             0.677 0.498730
## wealth.how.industry_Energy                              1.602 0.109365
## `wealth.how.industry_Hedge funds`                       0.514 0.607585
## wealth.how.industry_Media                               1.953 0.051021 .
## `wealth.how.industry_Mining and metals`                 1.646 0.099852 .
## `wealth.how.industry_Money Management`                  0.817 0.414297
## `wealth.how.industry_Non-consumer industrial`           1.736 0.082672 .
## wealth.how.industry_Other                               0.975 0.329676
## `wealth.how.industry_Private equity/leveraged buyout`   1.531 0.126036
## `wealth.how.industry_Real Estate`                       0.622 0.533801
## `wealth.how.industry_Technology-Computer`               1.949 0.051423 .
## `wealth.how.industry_Technology-Medical`                1.638 0.101621
## `wealth.how.industry_Venture Capital`                  -0.138 0.890643
## `wealth.how.inherited_4th generation`                   0.936 0.349253
## `wealth.how.inherited_5th generation or longer`        -0.432 0.666085
## wealth.how.inherited_father                             1.289 0.197441
## `wealth.how.inherited_not inherited`                   -0.156 0.876070
## `wealth.how.inherited_spouse/widow`                    -0.956 0.339342
## `wealth.type_founder non-finance`                       2.891 0.003888 **
## wealth.type_inherited                                   0.544 0.586215
## `wealth.type_privatized and resources`                  2.535 0.011337 *
## `wealth.type_self-made finance`                         0.524 0.600648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.482 on 1938 degrees of freedom
## Multiple R-squared:  0.09995,    Adjusted R-squared:  0.07766
## F-statistic: 4.484 on 48 and 1938 DF,  p-value: < 2.2e-16
```

```
### Pronalazak najboljih prediktora na sljedeći način: fittaj linearnu regresiju na svakom indikatoru po
### sortaj najbolje regressore po p vrijednosti

n = 10
filtered_col_names = c()
r_squares = c()
ps = c()
col_names=colnames(x)

for(i in 1:ncol(x)){

  col_name=col_names[i]
  model=lm(wealth ~ x[[col_name]]) # create model with a single regressor and predict wealth
  summary_model = summary(model)

  filtered_col_names <- append(filtered_col_names, col_name)
  r_squares <- append(r_squares, summary_model$r.squared)
  # Density, distribution function, quantile function and random generation for the F distribution with
  # a.k.a get P value from f statistics
  f <- summary_model$fstatistic
  ps <- append(ps,  pf(f[1], f[2], f[3], lower.tail=FALSE))
}

df_g_squares=data.frame(filtered_col_names, r_squares, ps)
head(df_g_squares, n=3)
```

```
##              filtered_col_names    r_squares          ps
## 1               company.founded 8.130424e-05 0.687912207
## 2 company.relationship_chairman 4.348546e-03 0.003273294
## 3 company.relationship_director 6.218927e-04 0.266527863
```

```
df_g_squares
```

```
##                                     filtered_col_names    r_squares
## 1                                      company.founded 8.130424e-05
## 2                        company.relationship_chairman 4.348546e-03
## 3                        company.relationship_director 6.218927e-04
## 4                         company.relationship_founder 4.031686e-03
## 5                        company.relationship_investor 1.402451e-03
## 6                           company.relationship_owner 1.185095e-02
## 7                       company.relationship_president 7.693116e-08
## 8                                 company.type_aquired 1.467008e-05
## 9                                     company.type_new 6.958026e-05
## 10                          company.type_privatization 1.516366e-03
## 11                             company.type_subsidiary 2.083085e-03
## 12                                    demographics.age 1.264222e-02
## 13                            demographics.gender_male 1.590454e-06
## 14                              location.region_Europe 1.613209e-03
## 15                       location.region_Latin America 4.530739e-03
## 16            location.region_Middle East/North Africa 9.431594e-04
## 17                       location.region_North America 4.616901e-03
## 18                          location.region_South Asia 2.264762e-03
## 19                  location.region_Sub-Saharan Africa 5.982409e-04
## 20                       wealth.how.category_Financial 6.676925e-06
## 21                     wealth.how.category_New Sectors 8.110268e-04
## 22              wealth.how.category_Non-Traded Sectors 3.157021e-03
## 23                wealth.how.category_Resource Related 1.663292e-03
## 24                  wealth.how.category_Traded Sectors 1.845987e-04
## 25                     wealth.how.industry_Constrution 1.398578e-03
## 26                        wealth.how.industry_Consumer 5.383965e-04
## 27           wealth.how.industry_Diversified financial 8.589656e-05
## 28                          wealth.how.industry_Energy 3.716918e-04
## 29                     wealth.how.industry_Hedge funds 1.342533e-03
## 30                           wealth.how.industry_Media 5.597208e-03
## 31               wealth.how.industry_Mining and metals 9.388670e-04
## 32                wealth.how.industry_Money Management 7.218781e-04
## 33         wealth.how.industry_Non-consumer industrial 4.916271e-04
## 34                           wealth.how.industry_Other 7.636769e-04
## 35 wealth.how.industry_Private equity/leveraged buyout 7.749904e-04
## 36                     wealth.how.industry_Real Estate 3.220077e-04
## 37             wealth.how.industry_Technology-Computer 1.049251e-05
## 38              wealth.how.industry_Technology-Medical 2.265278e-03
## 39                 wealth.how.industry_Venture Capital 9.310569e-04
## 40                 wealth.how.inherited_4th generation 2.140269e-03
## 41       wealth.how.inherited_5th generation or longer 8.037852e-05
## 42                         wealth.how.inherited_father 1.881069e-02
## 43                  wealth.how.inherited_not inherited 2.357067e-02
## 44                   wealth.how.inherited_spouse/widow 8.820890e-06
## 45                     wealth.type_founder non-finance 1.890736e-05
## 46                               wealth.type_inherited 2.355548e-02
## 47                wealth.type_privatized and resources 3.681178e-03
## 48                       wealth.type_self-made finance 6.974850e-03
##              ps
## 1  6.879122e-01
## 2  3.273294e-03
## 3  2.665279e-01
## 4  4.633888e-03
## 5  9.514366e-02
## 6  1.150257e-06
## 7  9.901416e-01
## 8  8.645184e-01
## 9  7.101904e-01
## 10 8.267516e-02
## 11 4.192585e-02
## 12 5.040078e-07
## 13 9.551979e-01
## 14 7.345843e-02
## 15 2.682592e-03
## 16 1.711795e-01
## 17 2.442123e-03
## 18 3.390440e-02
## 19 2.758214e-01
## 20 9.083576e-01
## 21 2.044727e-01
## 22 1.224519e-02
## 23 6.913027e-02
## 24 5.449896e-01
## 25 9.560201e-02
## 26 3.012312e-01
## 27 6.796953e-01
## 28 3.903795e-01
## 29 1.025104e-01
## 30 8.453764e-04
## 31 1.721573e-01
## 32 2.312617e-01
## 33 3.232196e-01
## 34 2.182112e-01
## 35 2.148315e-01
## 36 4.240270e-01
## 37 8.852636e-01
## 38 3.388406e-02
## 39 1.739529e-01
## 40 3.920606e-02
## 41 6.896011e-01
## 42 8.312478e-10
## 43 5.980060e-12
## 44 8.947415e-01
## 45 8.464061e-01
## 46 6.074959e-12
## 47 6.823784e-03
## 48 1.938178e-04
```

```
# sort (by minimal r_squares) and find top n predictors
df_top_predictors = df_g_squares[order(-df_g_squares$r_squares), ]

top_n_predictors_one_var_lin = as.vector(df_top_predictors$filtered_col_names)[1:n]
df_top_predictors
```

```
##                                     filtered_col_names    r_squares
## 43                  wealth.how.inherited_not inherited 2.357067e-02
## 46                               wealth.type_inherited 2.355548e-02
## 42                         wealth.how.inherited_father 1.881069e-02
## 12                                    demographics.age 1.264222e-02
## 6                           company.relationship_owner 1.185095e-02
## 48                       wealth.type_self-made finance 6.974850e-03
## 30                           wealth.how.industry_Media 5.597208e-03
## 17                       location.region_North America 4.616901e-03
## 15                       location.region_Latin America 4.530739e-03
## 2                        company.relationship_chairman 4.348546e-03
## 4                         company.relationship_founder 4.031686e-03
## 47                wealth.type_privatized and resources 3.681178e-03
## 22              wealth.how.category_Non-Traded Sectors 3.157021e-03
## 38              wealth.how.industry_Technology-Medical 2.265278e-03
## 18                          location.region_South Asia 2.264762e-03
## 40                 wealth.how.inherited_4th generation 2.140269e-03
## 11                             company.type_subsidiary 2.083085e-03
## 23                wealth.how.category_Resource Related 1.663292e-03
## 14                              location.region_Europe 1.613209e-03
## 10                          company.type_privatization 1.516366e-03
## 5                        company.relationship_investor 1.402451e-03
## 25                     wealth.how.industry_Constrution 1.398578e-03
## 29                     wealth.how.industry_Hedge funds 1.342533e-03
## 16            location.region_Middle East/North Africa 9.431594e-04
## 31               wealth.how.industry_Mining and metals 9.388670e-04
## 39                 wealth.how.industry_Venture Capital 9.310569e-04
## 21                     wealth.how.category_New Sectors 8.110268e-04
## 35 wealth.how.industry_Private equity/leveraged buyout 7.749904e-04
## 34                           wealth.how.industry_Other 7.636769e-04
## 32                wealth.how.industry_Money Management 7.218781e-04
## 3                        company.relationship_director 6.218927e-04
## 19                  location.region_Sub-Saharan Africa 5.982409e-04
## 26                        wealth.how.industry_Consumer 5.383965e-04
## 33         wealth.how.industry_Non-consumer industrial 4.916271e-04
## 28                          wealth.how.industry_Energy 3.716918e-04
## 36                     wealth.how.industry_Real Estate 3.220077e-04
## 24                  wealth.how.category_Traded Sectors 1.845987e-04
## 27            wealth.how.industry_Diversified financial 8.589656e-05
## 1                                       company.founded 8.130424e-05
## 41        wealth.how.inherited_5th generation or longer 8.037852e-05
## 9                                      company.type_new 6.958026e-05
## 45                      wealth.type_founder non-finance 1.890736e-05
## 8                                  company.type_aquired 1.467008e-05
## 37              wealth.how.industry_Technology-Computer 1.049251e-05
## 44                    wealth.how.inherited_spouse/widow 8.820890e-06
## 20                        wealth.how.category_Financial 6.676925e-06
## 13                             demographics.gender_male 1.590454e-06
## 7                        company.relationship_president 7.693116e-08
##               ps
## 43 5.980060e-12
## 46 6.074959e-12
## 42 8.312478e-10
## 12 5.040078e-07
## 6  1.150257e-06
## 48 1.938178e-04
## 30 8.453764e-04
## 17 2.442123e-03
## 15 2.682592e-03
## 2  3.273294e-03
## 4  4.633888e-03
## 47 6.823784e-03
## 22 1.224519e-02
## 38 3.388406e-02
## 18 3.390440e-02
## 40 3.920606e-02
## 11 4.192585e-02
## 23 6.913027e-02
## 14 7.345843e-02
## 10 8.267516e-02
## 5  9.514366e-02
## 25 9.560201e-02
## 29 1.025104e-01
## 16 1.711795e-01
## 31 1.721573e-01
## 39 1.739529e-01
## 21 2.044727e-01
## 35 2.148315e-01
## 34 2.182112e-01
## 32 2.312617e-01
## 3  2.665279e-01
## 19 2.758214e-01
## 26 3.012312e-01
## 33 3.232196e-01
## 28 3.903795e-01
## 36 4.240270e-01
## 24 5.449896e-01
## 27 6.796953e-01
## 1  6.879122e-01
## 41 6.896011e-01
## 9  7.101904e-01
## 45 8.464061e-01
## 8  8.645184e-01
## 37 8.852636e-01
## 44 8.947415e-01
## 20 9.083576e-01
## 13 9.551979e-01
## 7  9.901416e-01
```

```
# pronalazak najboljih regressora s ANOVA-om
# nađi P vrijednosti za svaki regressor
# mergaj regressore od prošlog koraka i ukolni duplikate
# dobivene regresore koristi za model

a <- anova(model_all_vars)
ps_a <- a$`Pr(>F)`
ps_a <- head(ps_a, -1) # anova returns NA for last element

ps_a_ord <- order(ps_a)
sorted_cols <- colnames(x)[order(colnames(x))]
top_predictors_anova <- sorted_cols[ps_a_ord][1:n]
cat ("Best ANOVA regressors:")
```

```
## Best ANOVA regressors:
```

```
top_predictors_anova
```

```
##  [1] "company.relationship_owner"
##  [2] "demographics.age"
##  [3] "company.relationship_founder"
##  [4] "wealth.how.inherited_father"
##  [5] "company.type_subsidiary"
##  [6] "location.region_Latin America"
##  [7] "company.relationship_chairman"
##  [8] "wealth.type_privatized and resources"
##  [9] "wealth.how.industry_Technology-Computer"
## [10] "wealth.type_founder non-finance"
```

```
top_predictors = c(top_predictors_anova, top_n_predictors_one_var_lin)
top_predictors <- top_predictors[!duplicated(top_predictors)]
cat ("\nTop predictors for a new model:")
```

```
##
## Top predictors for a new model:
```

```
top_predictors
```

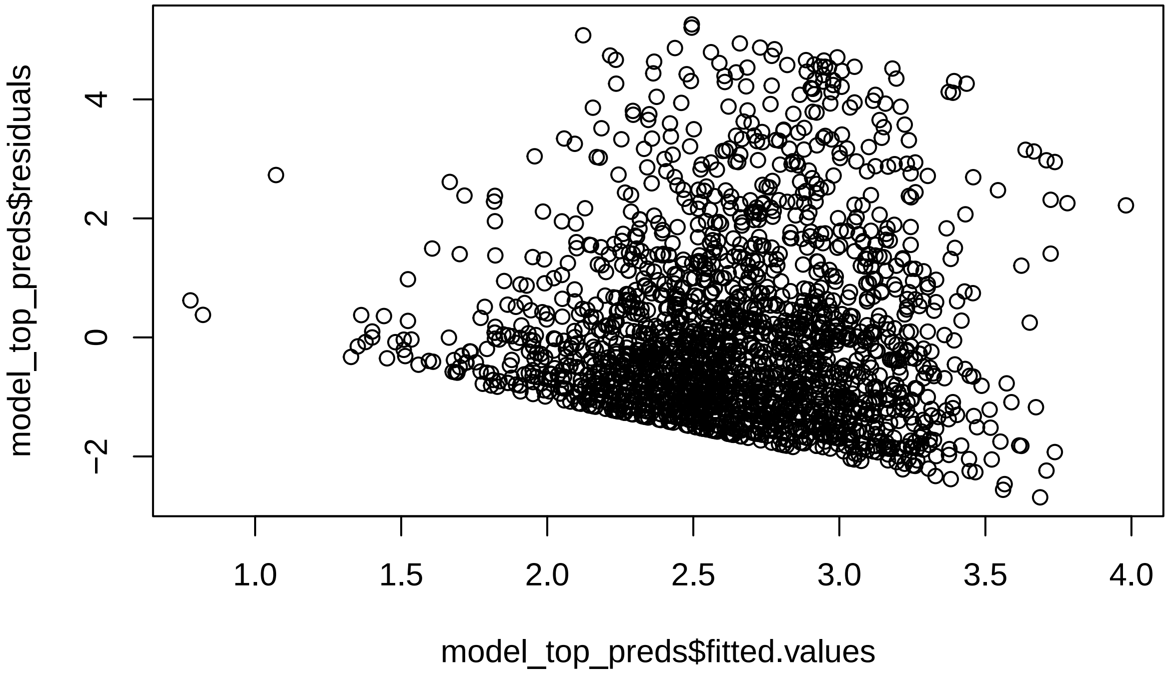
```
##  [1] "company.relationship_owner"
##  [2] "demographics.age"
##  [3] "company.relationship_founder"
##  [4] "wealth.how.inherited_father"
##  [5] "company.type_subsidiary"
##  [6] "location.region_Latin America"
##  [7] "company.relationship_chairman"
##  [8] "wealth.type_privatized and resources"
##  [9] "wealth.how.industry_Technology-Computer"
## [10] "wealth.type_founder non-finance"
## [11] "wealth.how.inherited_not inherited"
## [12] "wealth.type_inherited"
## [13] "wealth.type_self-made finance"
## [14] "wealth.how.industry_Media"
## [15] "location.region_North America"
```

```
model_top_preds <- lm(wealth ~ . , x[, top_predictors])
summary(model_top_preds)
```

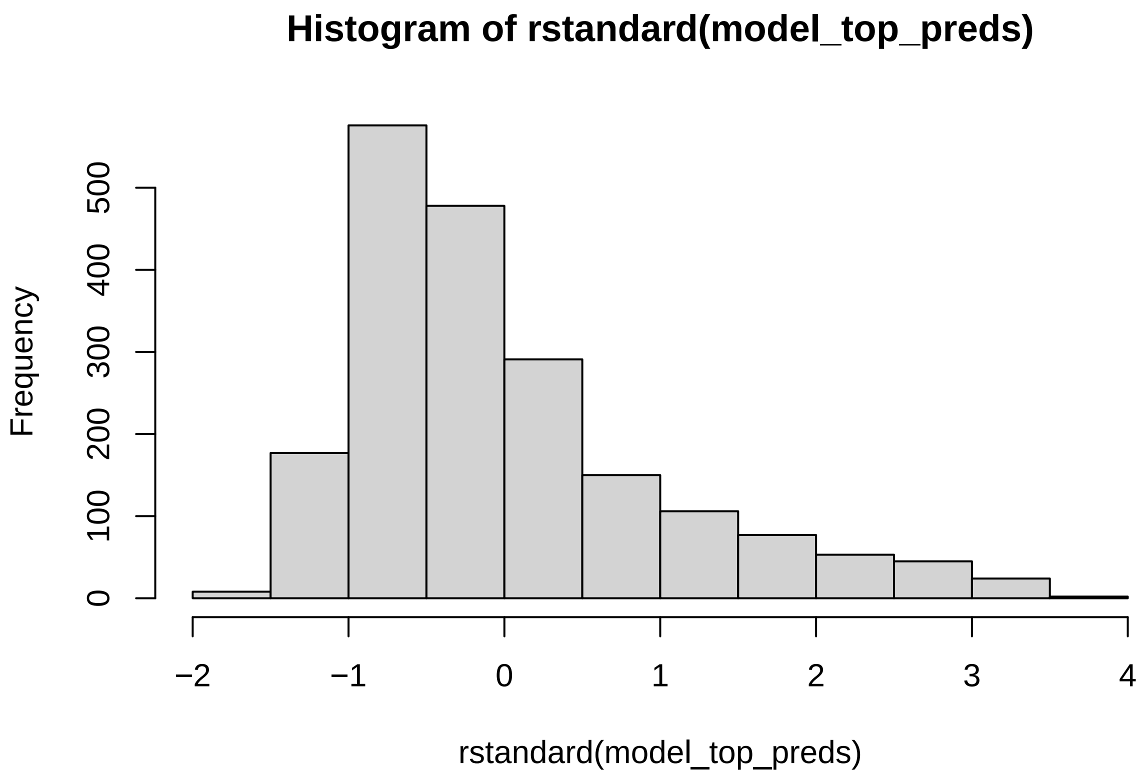
```
##
## Call:
## lm(formula = wealth ~ ., data = x[, top_predictors])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6874 -1.0582 -0.3905  0.6090  5.2597
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                1.64081    0.78452   2.091 0.036614
## company.relationship_owner                -0.64196    0.20445  -3.140 0.001715
## demographics.age                           1.23174    0.23070   5.339 1.04e-07
## company.relationship_founder              -0.22710    0.15197  -1.494 0.135228
## wealth.how.inherited_father                0.20164    0.11973   1.684 0.092325
## company.type_subsidiary                   -1.21450    0.61620  -1.971 0.048869
## `location.region_Latin America`           -0.49467    0.14103  -3.508 0.000462
## company.relationship_chairman             -0.28553    0.18475  -1.545 0.122394
## `wealth.type_privatized and resources`     0.42688    0.18596   2.296 0.021806
## `wealth.how.industry_Technology-Computer`  0.23898    0.13633   1.753 0.079776
## `wealth.type_founder non-finance`          0.47939    0.18709   2.562 0.010469
## `wealth.how.inherited_not inherited`      -0.09160    0.75409  -0.121 0.903330
## wealth.type_inherited                      0.47903    0.76232   0.628 0.529825
## `wealth.type_self-made finance`            0.30609    0.17643   1.735 0.082906
## wealth.how.industry_Media                  0.46363    0.12983   3.571 0.000364
## `location.region_North America`            0.07894    0.07254   1.088 0.276667
##
## (Intercept)                               *
## company.relationship_owner                **
## demographics.age                          ***
## company.relationship_founder
## wealth.how.inherited_father               .
## company.type_subsidiary                   *
## `location.region_Latin America`           ***
## company.relationship_chairman
## `wealth.type_privatized and resources`    *
## `wealth.how.industry_Technology-Computer` .
## `wealth.type_founder non-finance`         *
## `wealth.how.inherited_not inherited`
## wealth.type_inherited
## `wealth.type_self-made finance`           .
## wealth.how.industry_Media                 ***
## `location.region_North America`
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.495 on 1971 degrees of freedom
## Multiple R-squared:  0.06821,    Adjusted R-squared:  0.06112
## F-statistic: 9.619 on 15 and 1971 DF,  p-value: < 2.2e-16
```

```
# micanjem nekih od ovih regresora se povećava Adjusted R-squared
```

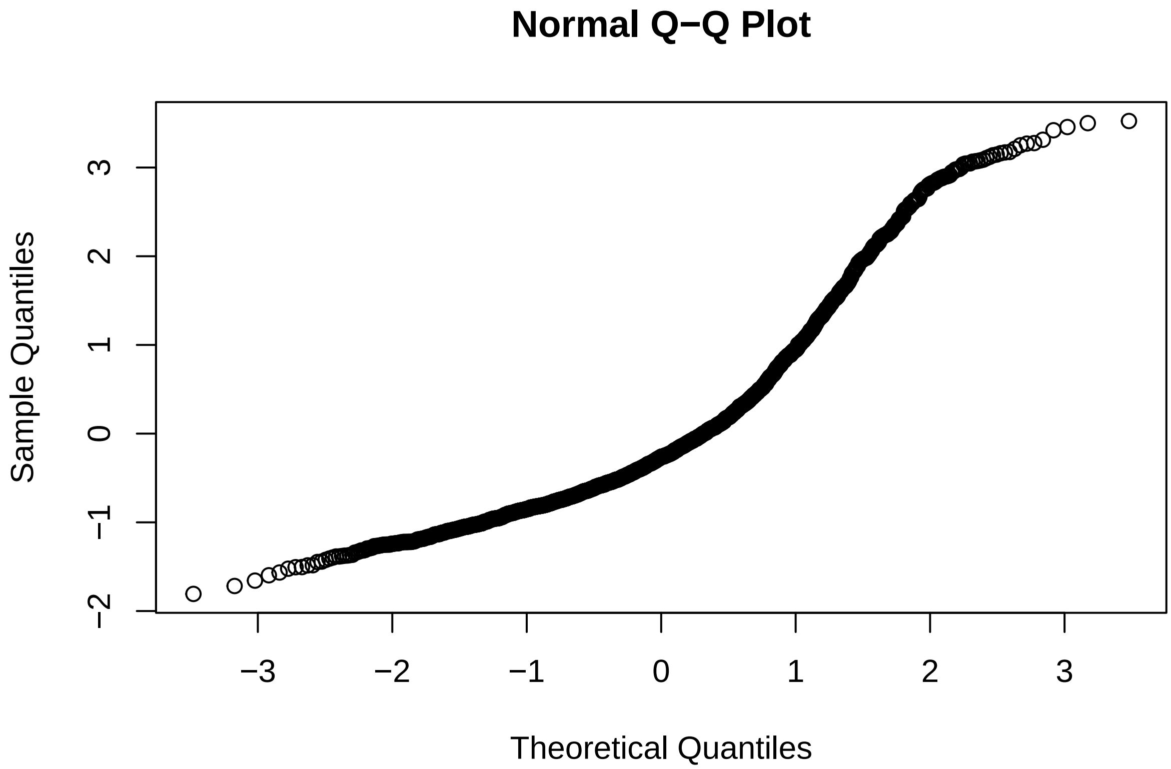
```
require(nortest)

# reziduali u ovisnosti o procjenama modela
plot(model_top_preds$fitted.values, model_top_preds$residuals)
```



```
# provjera normalnosti reziduala
hist(rstandard(model_top_preds))
```



```
qqnorm(rstandard(model_top_preds))
```



```
ks.test(rstandard(model_top_preds),'pnorm')
## Warning in ks.test(rstandard(model_top_preds), "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(model_top_preds)
## D = 0.12709, p-value < 2.2e-16
## alternative hypothesis: two-sided
lillie.test(rstandard(model_top_preds))
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(model_top_preds)
## D = 0.12731, p-value < 2.2e-16
```

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.

- 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

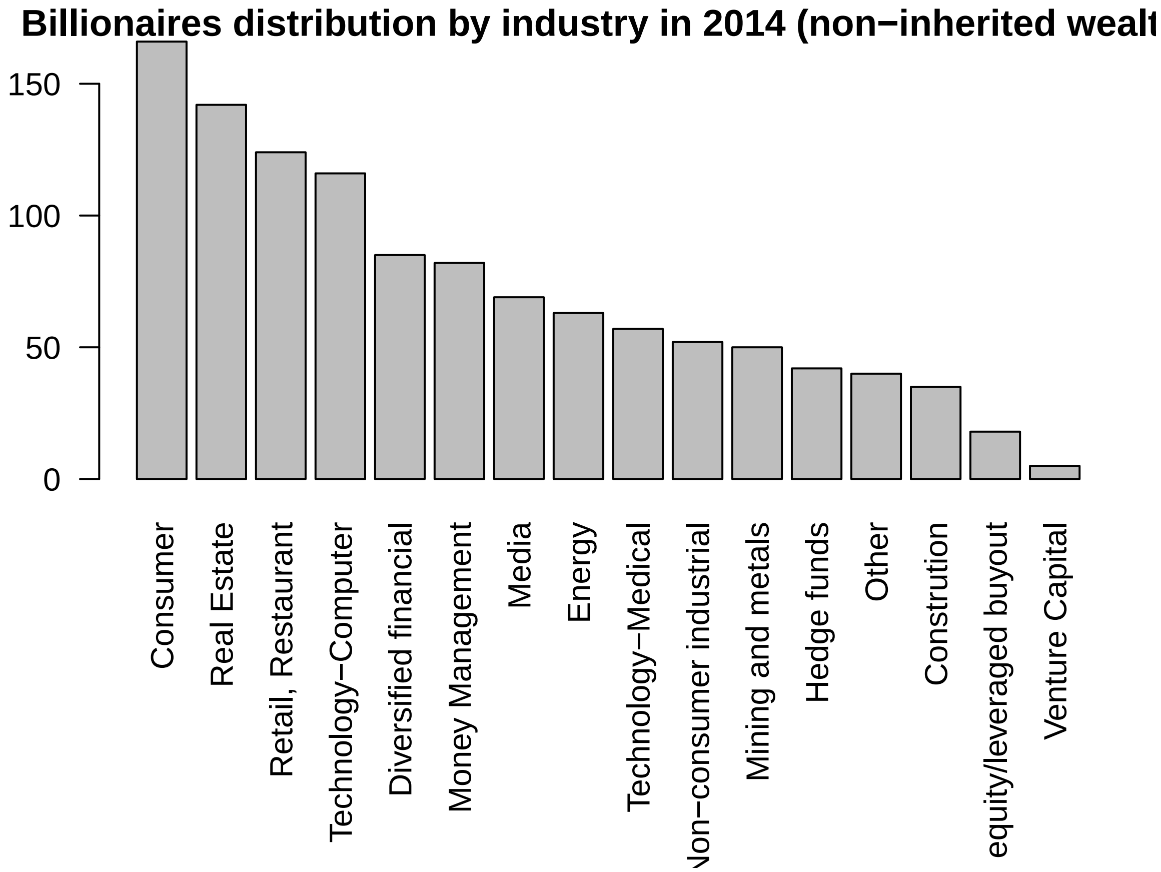
```
#
non_inherited_2014 <- non_inherited[non_inherited$year == 2014,]
non_inherited_2001 <- non_inherited[non_inherited$year == 2001,]
non_inherited_2014_new = bill_data[FALSE,]
```

```
## tracemem[0x60000272e4c0 -> 0x60000272af40]: lapply tbl_subset_row [.tbl_df [ eval eval withVisible wi
```

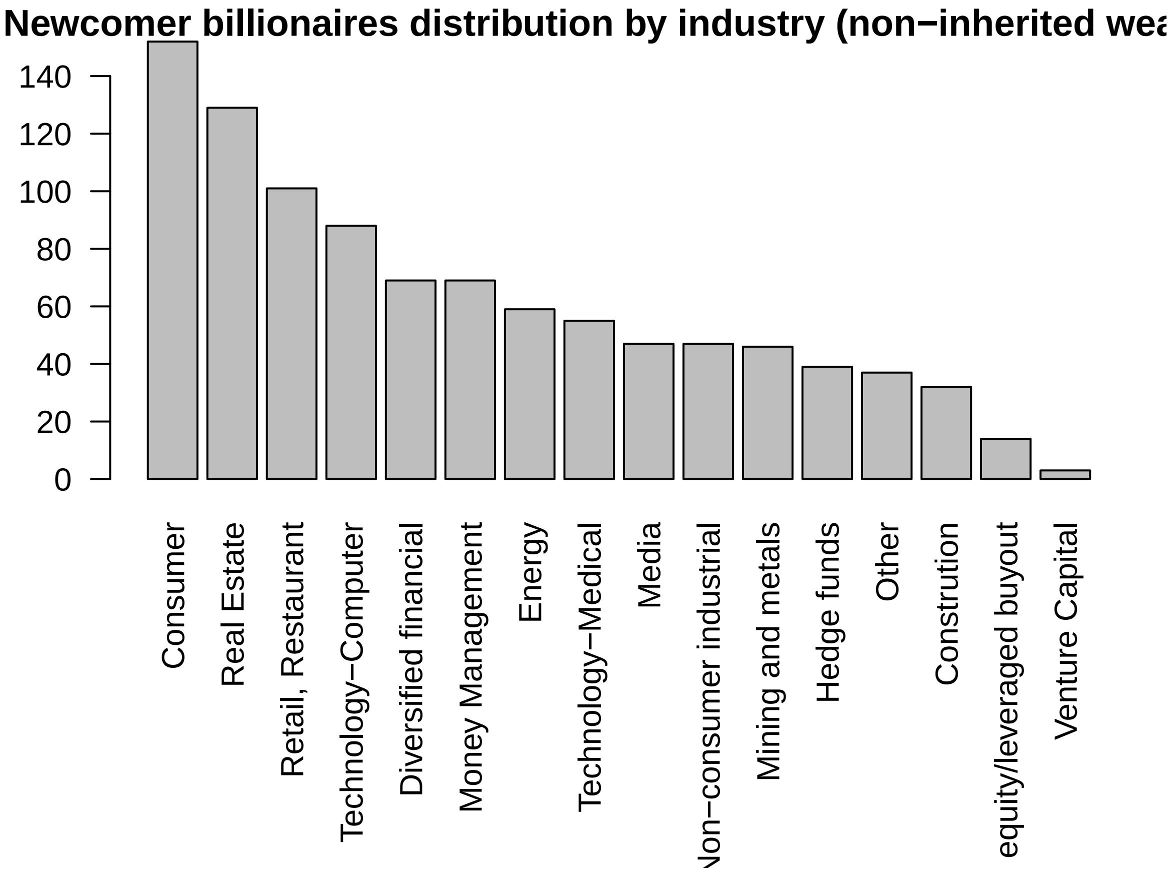
```
non_inherited_2001_old = bill_data[FALSE,]
```

```
## tracemem[0x60000272e4c0 -> 0x60000272b2c0]: lapply tbl_subset_row [.tbl_df [ eval eval withVisible wi
```

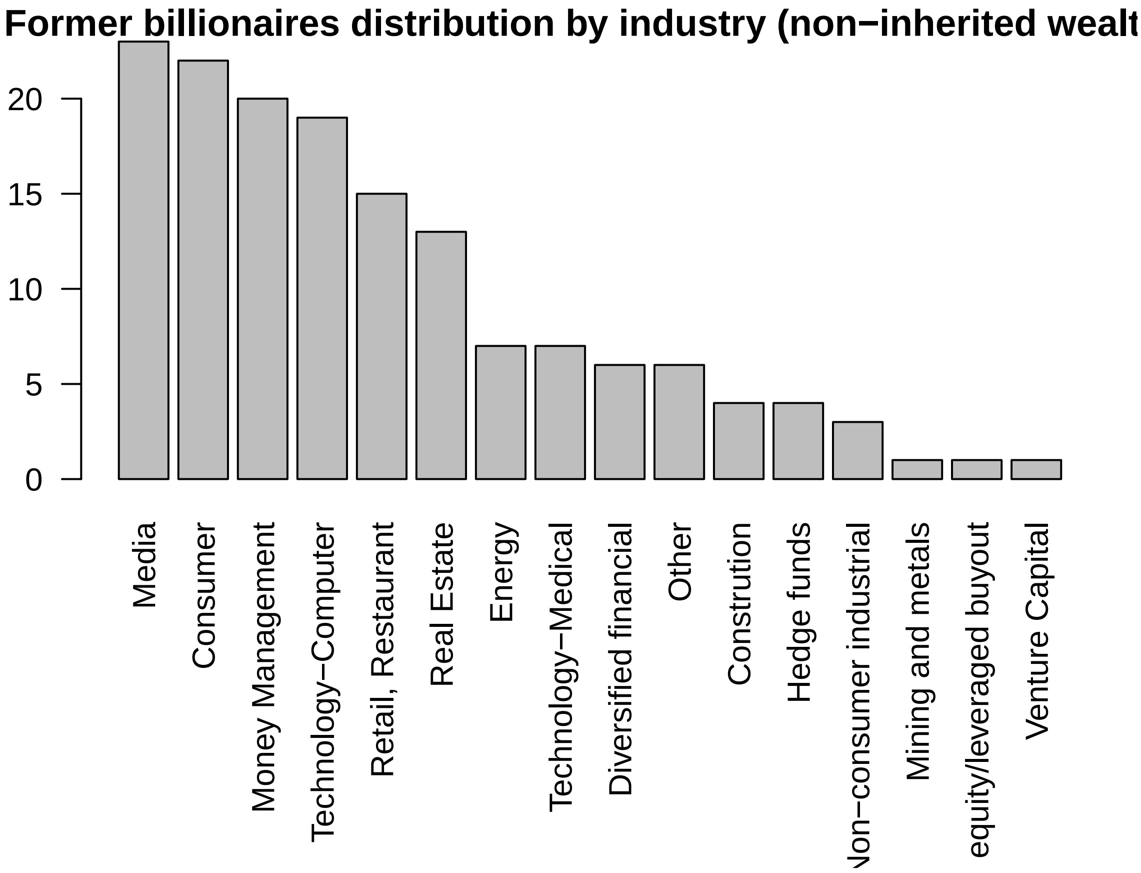
```
# selekcija novonastalih milijardera iz 2014. koji nisu bili na prethodnoj listi iz 2001.
for(i in 1:nrow(non_inherited_2014)) {
  r <- non_inherited_2014[i,]
  if(sum(str_detect(non_inherited_2001$name, r[[1]])) == 0) {
    non_inherited_2014_new <- rbind(non_inherited_2014_new, non_inherited_2014[i,])
  }
}

# selekcija milijardera iz 2001. koji nisu na listi iz 2014.
for(i in 1:nrow(non_inherited_2001)) {
  r <- non_inherited_2001[i,]
  if(sum(str_detect(non_inherited_2014$name, r[[1]])) == 0) {
    non_inherited_2001_old <- rbind(non_inherited_2001_old, non_inherited_2001[i,])
  }
}

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 in 2014 (non-inherited wealth)",
        las = 2)
```



```
barplot(sort(table(subset(non_inherited_2014_new$wealth.industry, non_inherited_2014_new$wealth.how
        main = "Newcomer billionaires distribution by industry (non-inherited wealth)",
        las = 2)
```



```
barplot(sort(table(subset(non_inherited_2001_old$wealth.how.industry, non_inherited_2001_old$wealth.how
        main = "Former billionaires distribution by industry (non-inherited wealth)",
        las = 2)
```



Former billionaires distribution by industry (non−inherited wealt