SAP - projekt - Milijarderi

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 bogatstvo 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?

Dodatno pitanje:

5. Jesu li muškarci milijarderi statistički značajno bogatiji od žena milijardera?

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
```

Svaki milijarder(2614) u danim podatcima sadrži 21 atribut koji ga opisuje. Neki od njih su: broj godina, spol, državljanstvo, porijeklo bogatstva, struka, vrijednost imovine, itd.

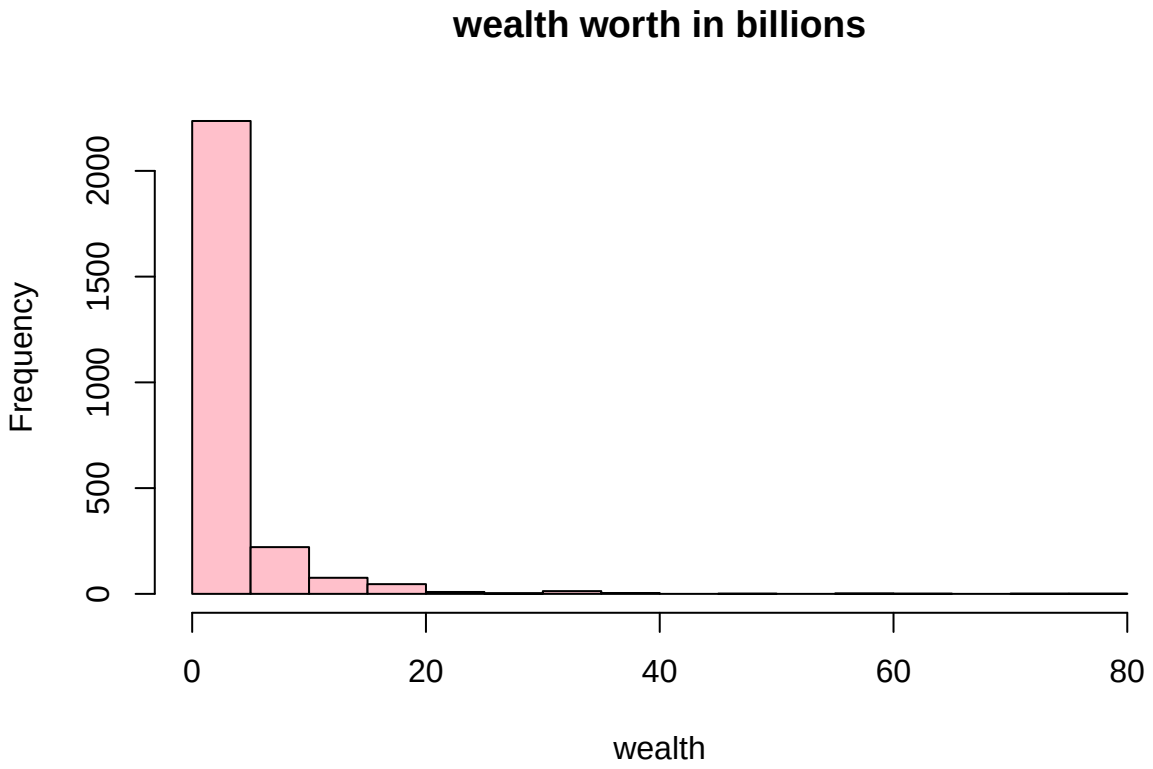
```
for (col_name in names(bill_data)){
  if (sum(is.na(bill_data[,col_name])) > 0){
    cat('Ukupno nedostajućih vrijednosti za varijablu'
        ,col_name, ': ', sum(is.na(bill_data[,col_name])),'\n')
  }
}
```

```
## 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
```

Naš dataset sastoji se od character i numeric varijabli. Prvo promotrimo numeričku varijablu wealth.

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

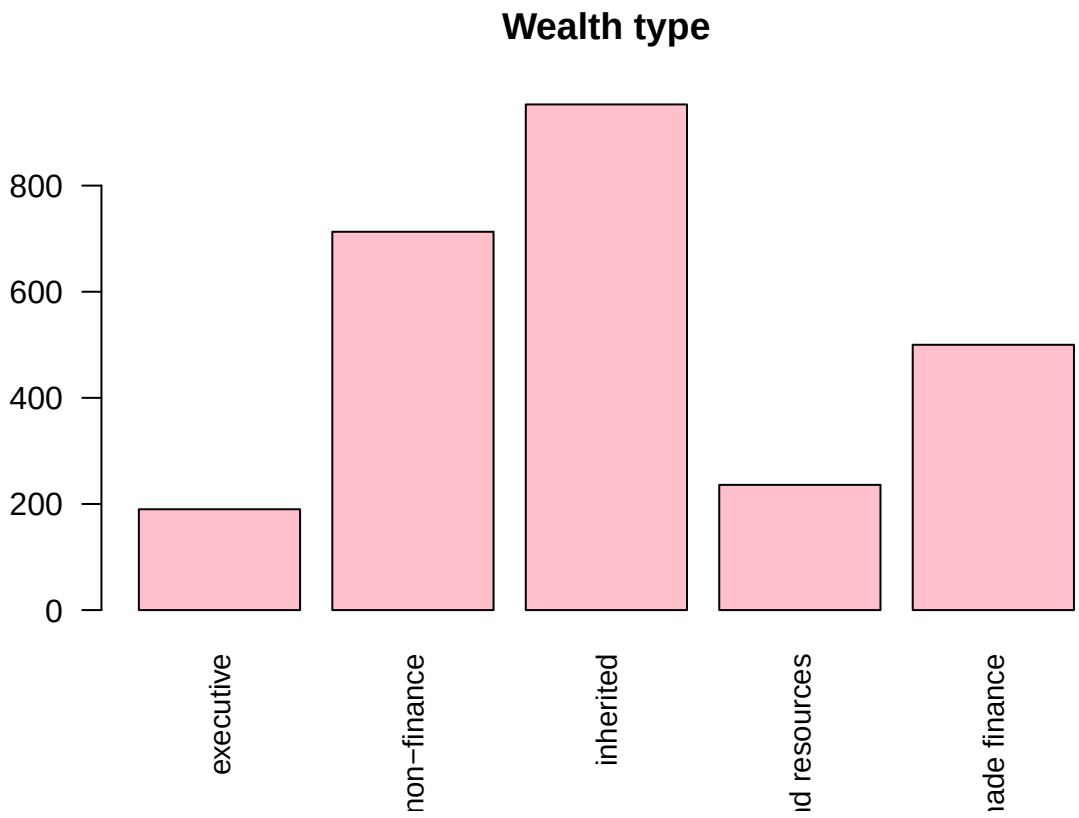



```
summary(bill_data$`wealth.worth in billions`)
```

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

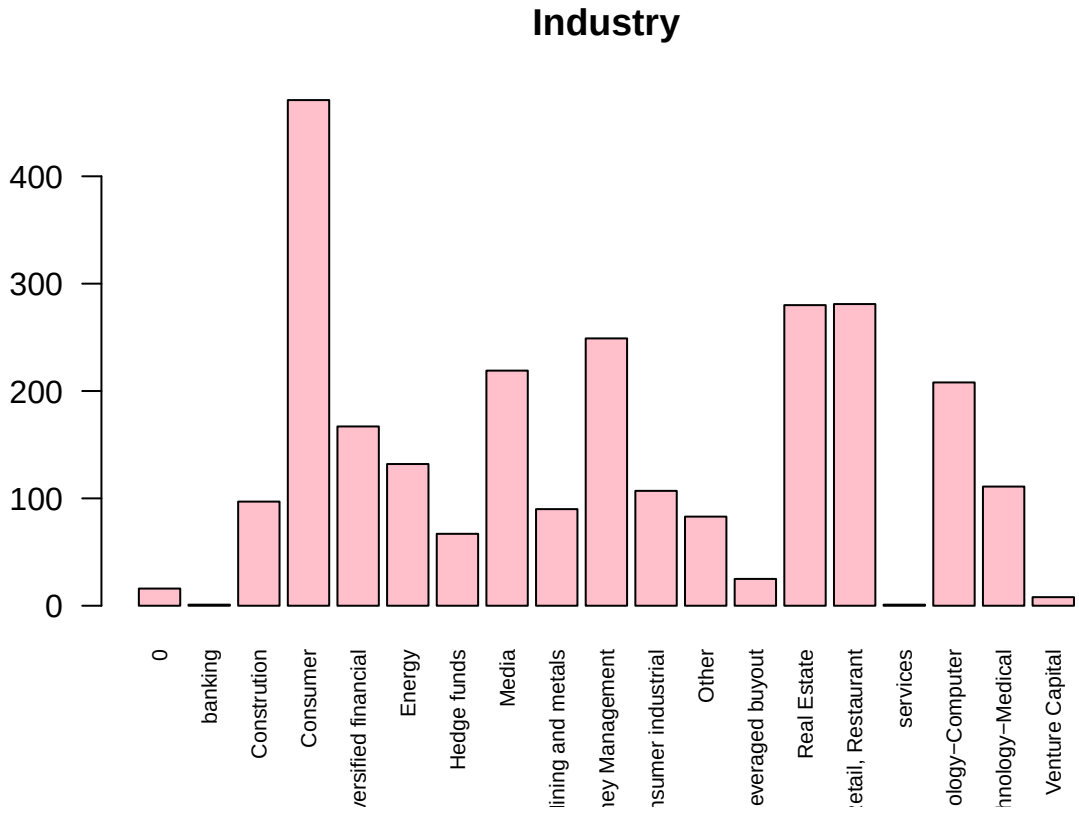
Ovaj histogram nam prikazuje distribuciju bogatstva. Možemo vidjeti da se "wealth" ne ravna po normalnoj distribuciji i da je jako malo onih s velikom količinom bogatstva. To nam pokazuju i rezultati koje smo dobili - srednja vrijednost i medijan.

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



Na ovom grafu vidimo koji tip bogatstva je najzastupljeniji. Najviše ima onih koji su naslijedili bogatstvo.

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



Na ovom grafu vidimo koje industrije su najzastupljenije kod milijardera. Najviše je onih u potrošačkoj industriji (consumer), a slijede ih oni koji se bave nekretninama, maloprodajom i restoranima.

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

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

Ovdje vidimo da muškarci značajno prevladavaju u broju milijardera naspram žena i supružnika.

```
median(bill_data$demographics.age)
```

```
## [1] 59
```

```
mean(bill_data$demographics.age)
```

```
## [1] 53.34124
```

Ako idemo proučiti milijardere po godinama, možemo vidjeti da je izračunata srednja vrijednost za varijablu starost 53 godine. Medijan iznosi 59 godina, odnosno kada bi ih poredali od najmlađeg do najstarijeg, vrijednost u sredini bila bi 59 godina. To znači da je više onih milijardera koji su stariji, odnosno u drugoj polovici života.

Pitanja

1. Ima li neki kontinent statistički značajno više milijardi?

Za početak želimo vidjeti je li svim milijarderima u našem datasetu dodijeljen kontinent. S obzirom na to da kontinent kao varijabla ne postoji, koristit ćemo regiju (location.region). Sada želimo izlistati sve regije koje postoje u datasetu.

```
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"
```

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
```

S obzirom na to da imamo regiju Middle East/North Africa trebamo ih rastaviti na kontinente kojima pripadaju (Azija i Afrika). Prvo želimo vidjeti koje sve države postoje u toj regiji u našem datasetu pomoću

državljanstva. Također imamo jednu državu čija regija ima vrijednost 0, a država je Bermuda. Dakle, nju ćemo kasnije u kodu svrstati pod kontinent Sjeverna Amerika.

```
bill_data$location.citizenship[bill_data$location.region == "Middle East/North Africa"]
bill_data$location.citizenship[bill_data$location.region == "0"]
```

Sada možemo združiti podatke ovisno o kontinentu.

Kopirajmo najprije podatke u novi data.frame (bill_data_copy).

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

```
## [1] FALSE
```

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

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

# Združimo 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";
}

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

  for (column_name in c("Bermuda")){
  bill_data_copy$location.region[bill_data_copy$location.citizenship == column_name] = "North America";
}
}

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

# Združimo 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";
}

#Združimo Australiju
for (column_name in c("Australia")){
  bill_data_copy$location.region[bill_data_copy$location.citizenship == column_name] = "Australia";
```

```
}


bill_data_copy
```

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

```
##
##        Africa          Asia     Australia        Europe North America
##            43           699            33           697           960
## South America
##           182
```

Sada kad smo završili s pripremom podataka za ovaj zadatak, možemo započeti sa statističkim testovima. ANOVA je parametarski test kojim se uspoređuju srednje vrijednosti više uzoraka te se na temelju F-testa donosi zaključak o postojanju značajnih razlika između tih srednjih vrijednosti. Na taj se način analizira utjecaj jedne ili više nezavisnih varijabli na jednu numeričku kontinuiranu (zavisnu) varijablu. U ovom slučaju razmatrat ćemo location.region kao varijablu koja određuje grupe (populacije) i wealth kao zavisnu varijablu. Pretpostavke jednofaktorske ANOVA-e su:

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

Nezavisnost uzoraka je zadovoljena s obzirom na to da jedna osoba ne potječe s više kontinenata.

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.

Provjeru normalnosti za svaku pojedinu grupu napravit ćemo Lillieforsovom inačicom KS testa.

Pretpostavke Lillieforsevog testa:

$$H_0: \textit{ podaci se ravnaju po normalnoj distribuciji}$$
$$H_1: \textit{ podaci se ne ravnaju po normalnoj distribuciji}$$

```
# logaritmirali smo wealth kako bi dobili ljepšu distribuciju na grafovima

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 = 0.12087, p-value < 2.2e-16
```

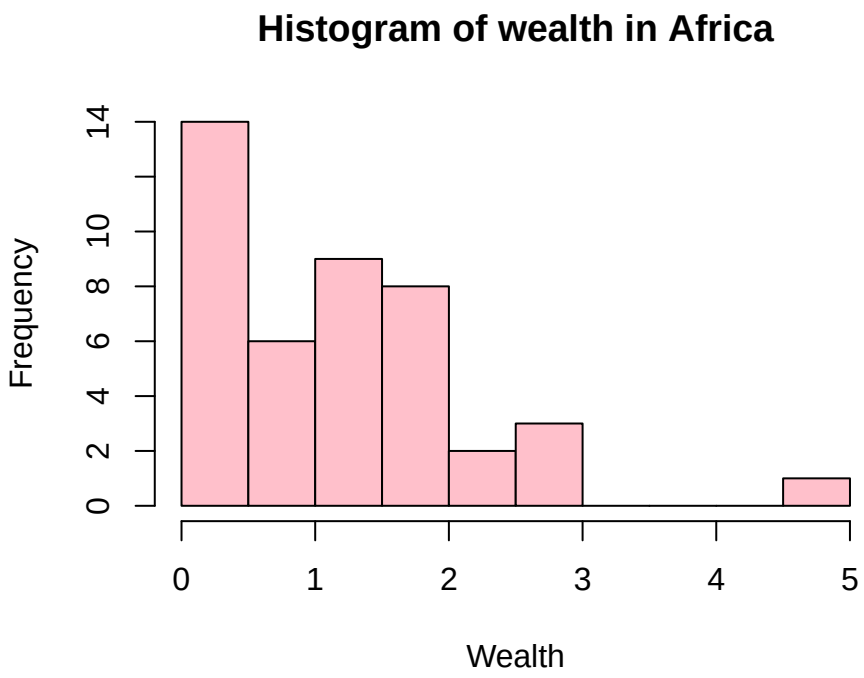
```
lillie.test(wealth[bill_data_copy$location.region=='Asia'])
```

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

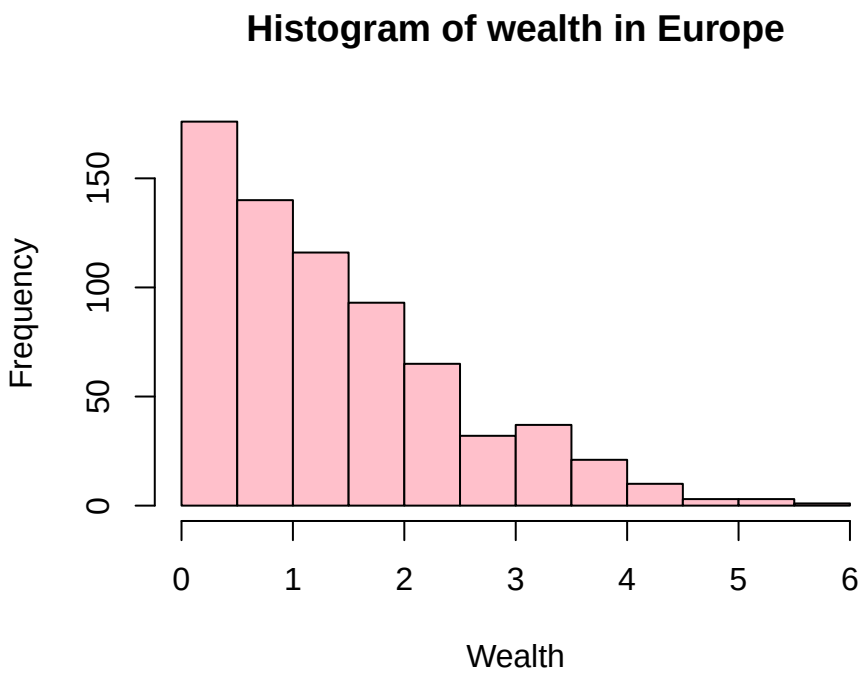
```
lillie.test(wealth[bill_data_copy$location.region=='Australia'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wealth[bill_data_copy$location.region == "Australia"]
## D = 0.14485, p-value = 0.07652
```

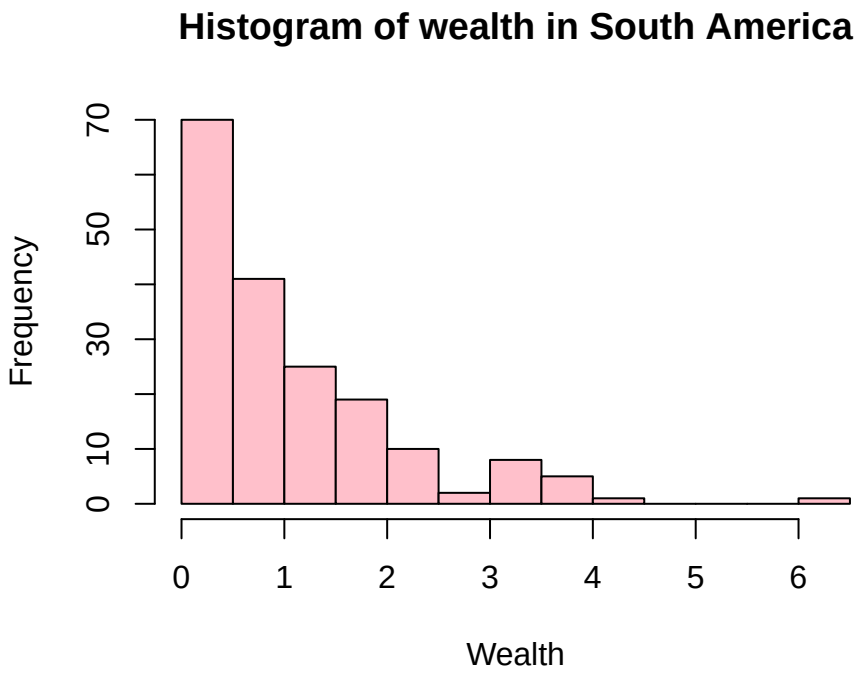
```
hist(wealth[bill_data_copy$location.region=='Africa'], main = "Histogram of wealth in Africa", xlab="Wea
```



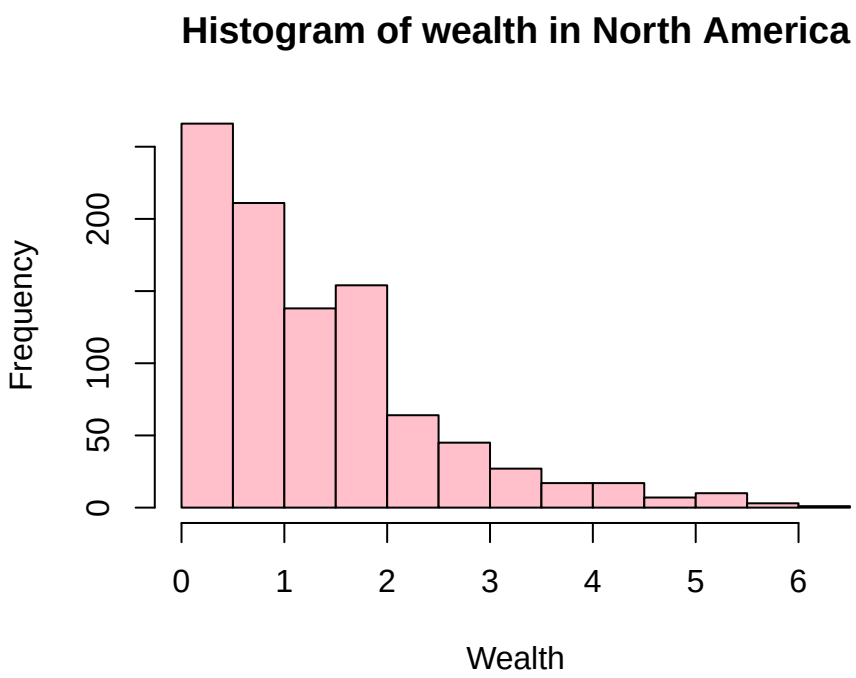

```
hist(wealth[bill_data_copy$location.region=='Europe'], main = "Histogram of wealth in Europe", xlab="Wea
```




```
hist(wealth[bill_data_copy$location.region=='South America'], main = "Histogram of wealth in South Ameri
```

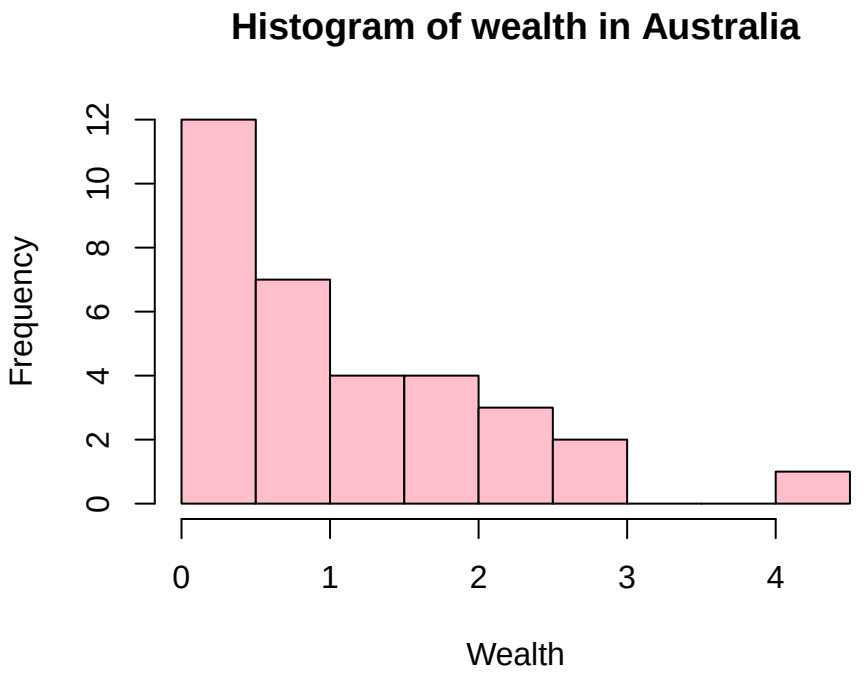



```
hist(wealth[bill_data_copy$location.region=='North America'], main = "Histogram of wealth in North Amer
```




```
hist(wealth[bill_data_copy$location.region=='Asia'], main = "Histogram of wealth in Asia", xlab="Wealth
```


```
hist(wealth[bill_data_copy$location.region=='Australia'], main = "Histogram of wealth in Australia", xla
```



Po rezultatima testa za normalnost (p vrijednosti manje od 0.05) te dobivenim histogramima vidimo da nam normalnost i nije zadovoljena. Nastavit ćemo s provjerom homogenosti varijanci Bartlettovim testom. Njegove pretpostavke su:

$$H_0 : \textit{ varijance su jednake}$$
$$H_1 : \textit{ varijance se razlikuju}$$

```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom

bartlett.test(wealth ~ bill_data_copy$location.region )
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  wealth by bill_data_copy$location.region
## Bartlett's K-squared = 20.71, df = 5, p-value = 0.0009188
```

```
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] 1.27337
```

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

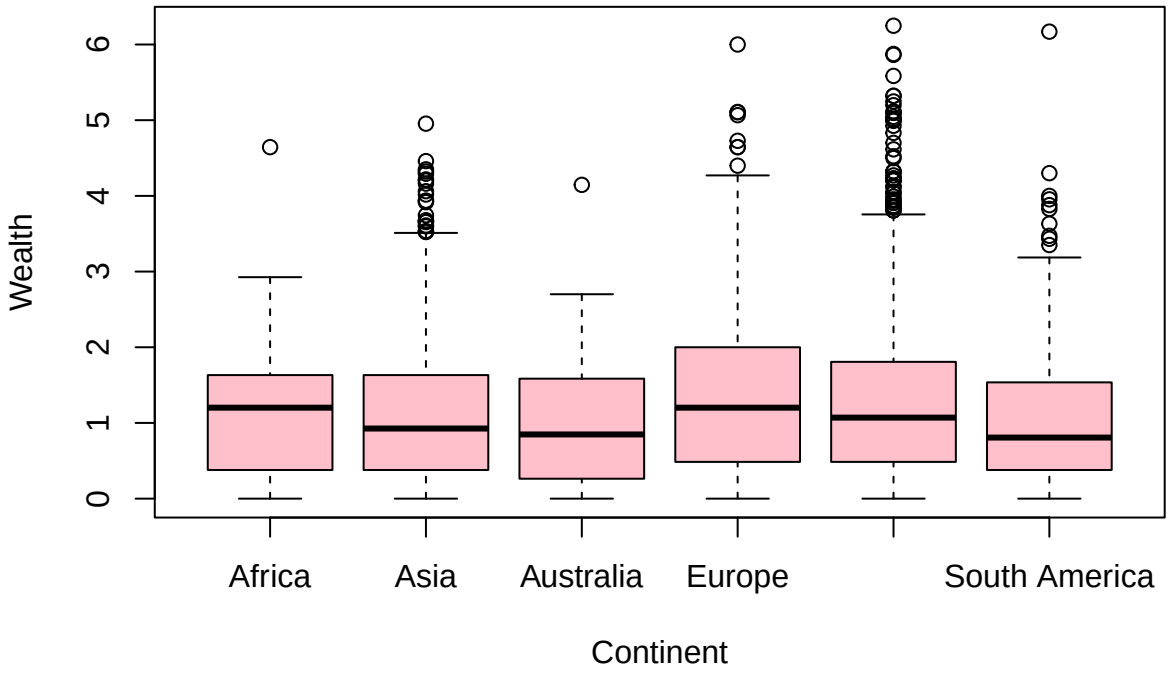
```
## [1] 1.076448
```

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

```
## [1] 0.9660385
```

S obzirom na to da na temelju ovih rezultata($p=0.0009 < 0.05$) odbacujemo nultu hipotezu da su nam varijance jednake, ne možemo koristiti ANOVA test. Prvo ćemo prikazati na boxplot grafu ovisnost varijable bogatstva o kontinentu kako bi grafički mogli interpretirati rezultat, a zatim ćemo utvrditi sigurnost rezultata koristiti neparametarski Kruskal - Wallis test kao alternativu jednofaktorskoj ANOVA-i.

```
# Grafički prikaz podataka
boxplot(wealth ~ bill_data_copy$location.region, xlab="Continent", ylab="Wealth", col= "pink")
```



Izračunajmo sada srednje vrijednosti bogatstva za svaki kontinent.

```r
mean_all= mean(bill_data_copy$wealth.worth.in.billions)
mean_all
```

```
## [1] 3.531943
```

```r
mean_by_continent <- bill_data_copy %>%
  group_by(location.region) %>%
  summarize(mean_continent = mean(wealth.worth.in.billions))
mean_by_continent
```

```
## # A tibble: 6 x 2
##   location.region mean_continent
##   <chr>                    <dbl>
## 1 Africa                    3.06
## 2 Asia                      2.94
## 3 Australia                 2.82
## 4 Europe                    3.80
## 5 North America             3.89
## 6 South America             3.17
```

Razmatrajući graf i numeričke rezultate koje smo dobili utvrdili smo da ima manje razlike između srednjih vrijednosti varijable wealth podijeljene po kontinentima.Sada idemo vidjeti je li ta razlika statistički značajna. Ovaj test služi za testiranje jednakosti srednjih vrijednosti u jednofaktorskoj analizi varijance. Pretpostavke Kruskal -Wallis testa :

H_0 : *nema razlike među populacijama*

H_1 : *postoji razlika među populacijama* (*barem dvije se razlikuju*)

```r
# Alternativa ANOVI - Kruskal - Wallis test

kruskal.test(bill_data_copy$wealth.worth.in.billions ~ bill_data_copy$location.region)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  bill_data_copy$wealth.worth.in.billions by bill_data_copy$location.region
## Kruskal-Wallis chi-squared = 33.298, df = 5, p-value = 3.284e-06
```

Kako je p-vrijednost manja od nivoa značajnosti od 0.05, možemo zaključiti da ima statistički značajne razlike između milijardi dijeljenim po kontinentima (među barem 2 kontinenta). Dakle, pomoću našeg izračuna srednjih vrijednosti i grafa možemo utvrditi da Sjeverna Amerika prednjači u količini bogatstva, a odmah iza nje nalazi se Europa.

2. Jesu li milijarderi koji su naslijedili bogatstvo 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 naslijedili bogatstvo) i non_inherited (oni koji nisu naslijedili bogatstvo).

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

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

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

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

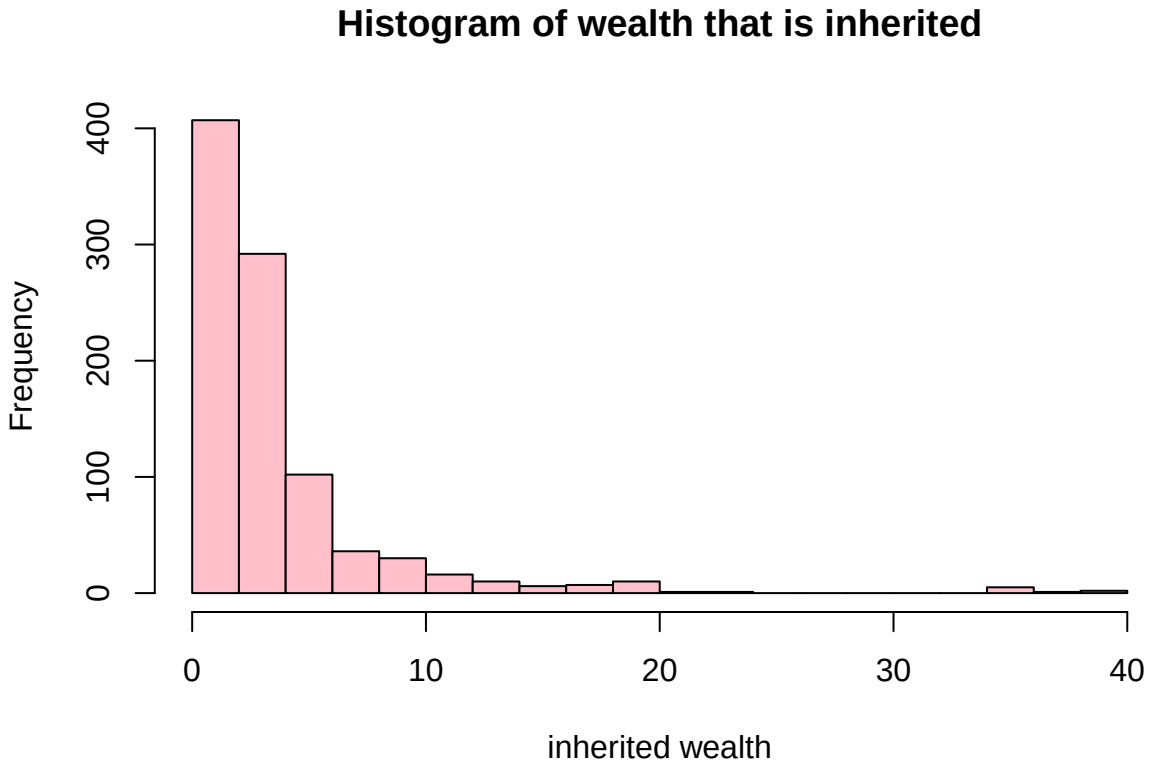
```
non_inherited_mean = mean(non_inherited$`wealth.worth in billions`)
print(non_inherited_mean)
```

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

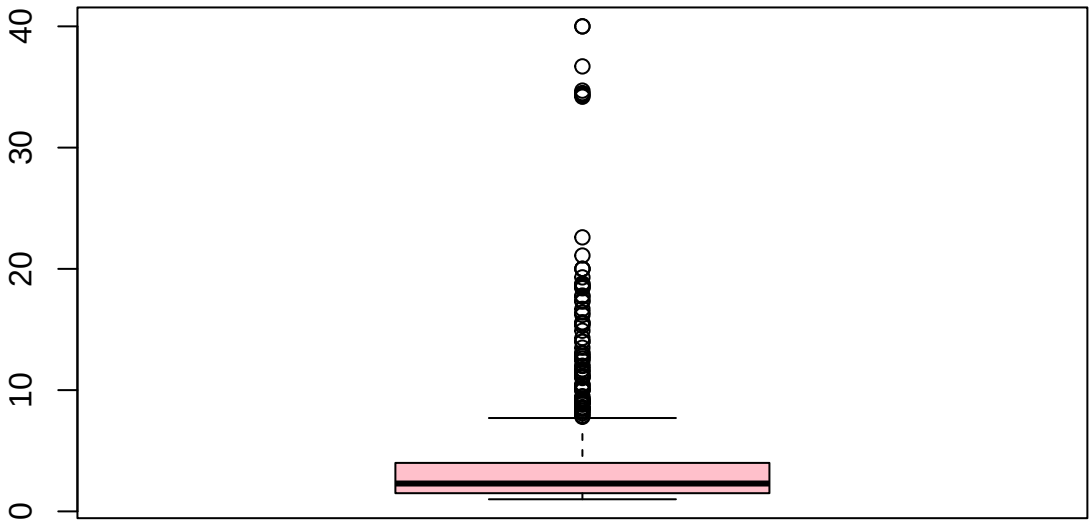
Na temelju male razlike u srednjim vrijednostima, ne postoje indikacije da su milijarderi koji su naslijedili bogatstvo statistički značajno bogatiji od onih koji nisu. No, navedeno je potrebno provjeriti.

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

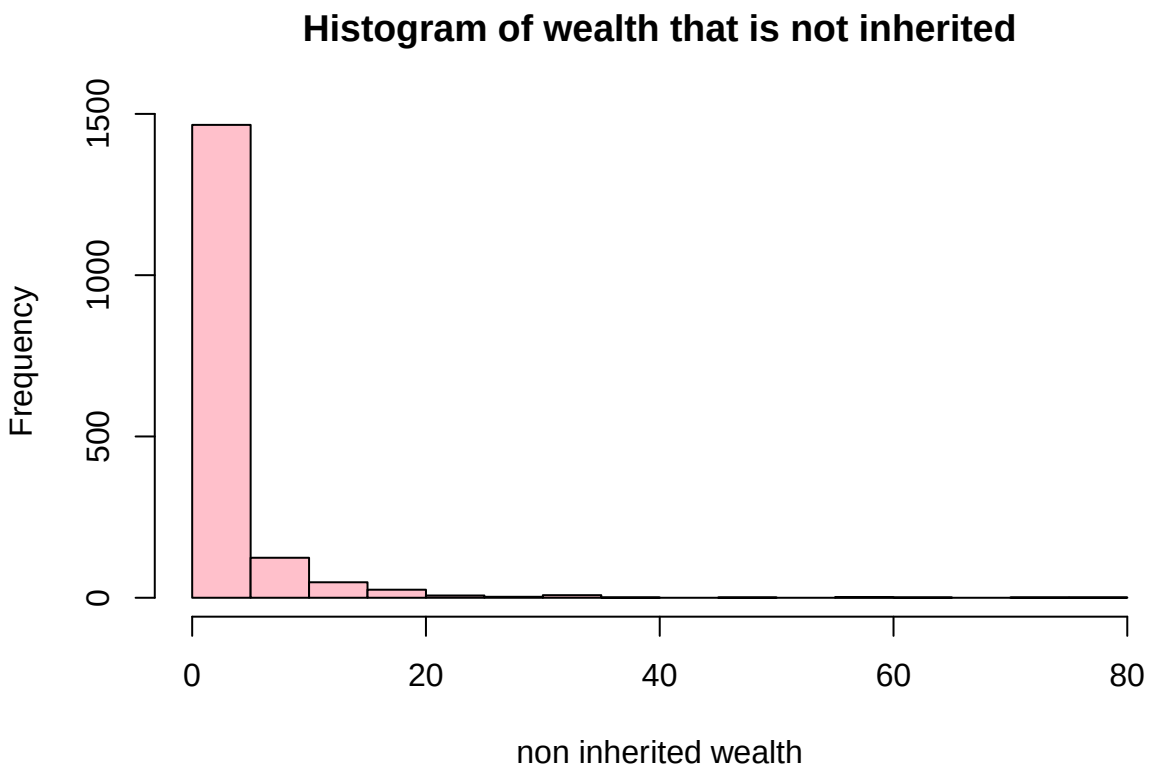
```
hist(inherited$`wealth.worth in billions`, breaks = 20, main = "Histogram of wealth that is inherited",
```



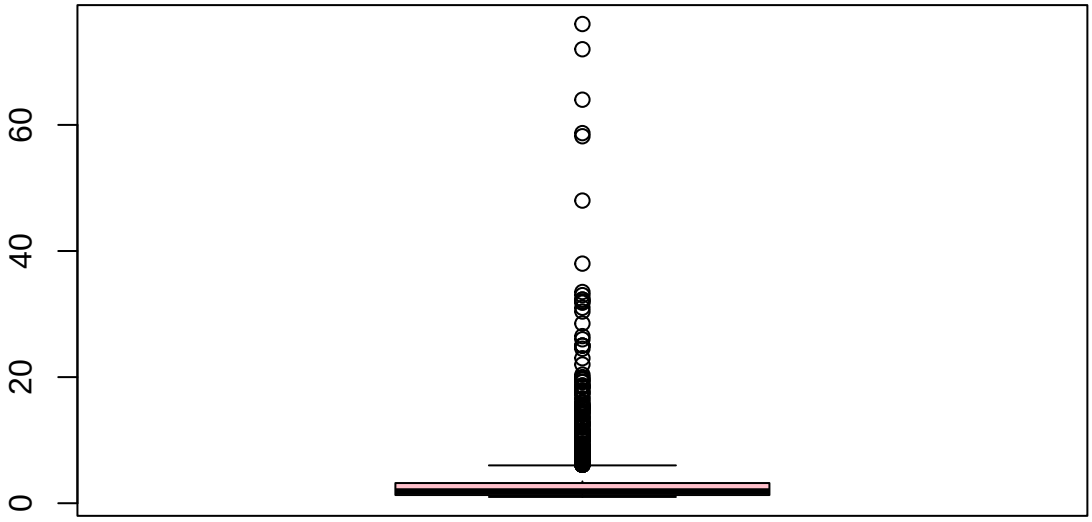

```
boxplot(inherited$`wealth.worth in billions`, col = "pink")
```



```
hist(non_inherited$`wealth.worth in billions`, breaks = 20, main = "Histogram of wealth that is not inh
```



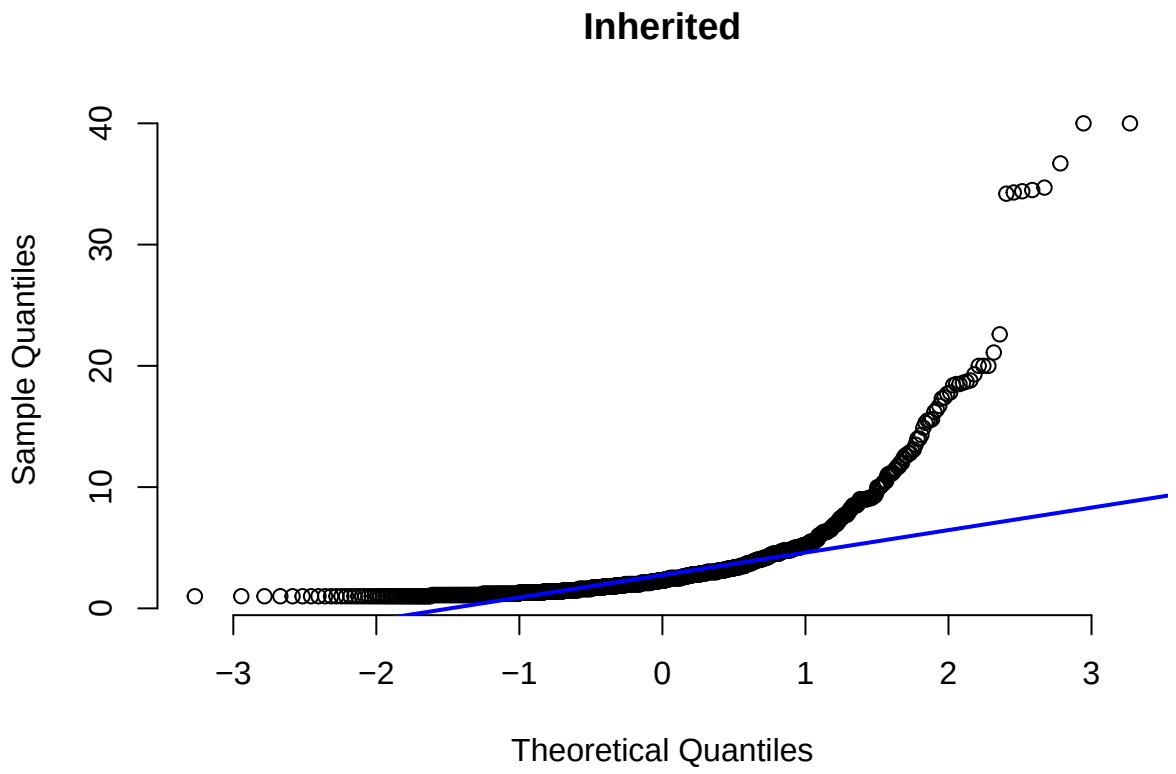

```
boxplot(non_inherited$`wealth.worth in billions`, col ="pink")
```



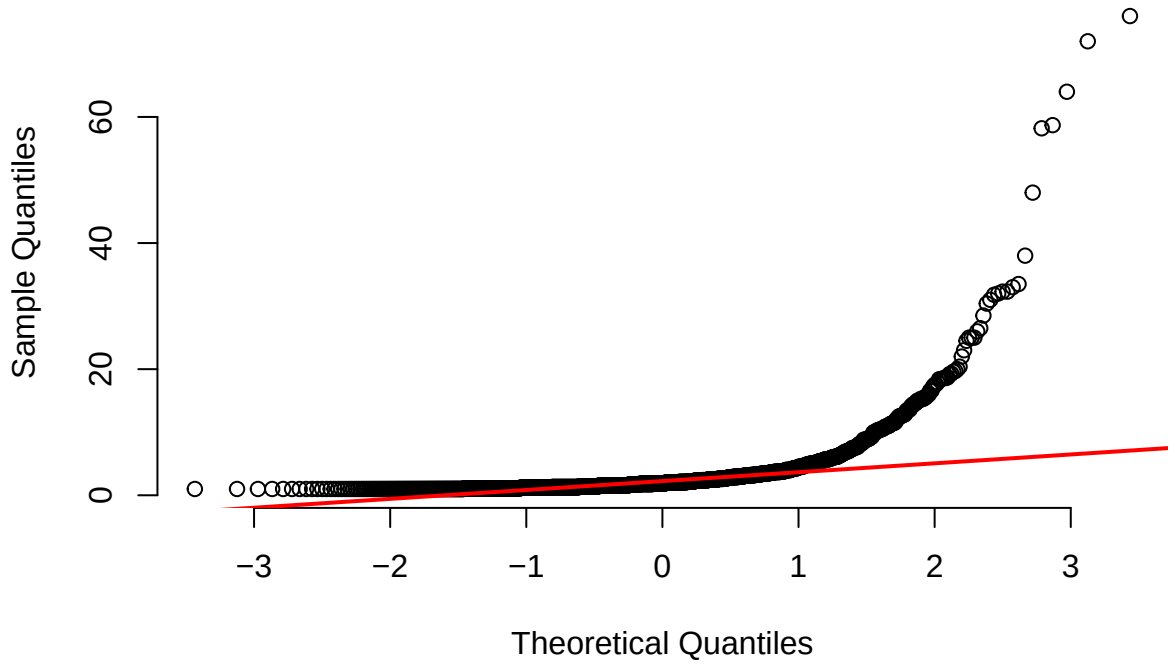
Iz prikazane vizualizacije uočavamo kako se podaci ne ravnaju po normalnoj distribuciji.

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")
```

```
##
##  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")
```

```
##
##  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 naslijediti i nenaslijediti 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 naslijedili bogatstvo nisu statistički značajno bogatiji od onih koji nisu.

3. Možete li iz danih varijabli predvidjeti njihovo bogatstvo?

Cilj ovog pitanja je provjeriti postoji li statistički značajna veza između više ulaznih varijabli (regresora) i izlazne varijable (reakcije, `wealth.worth in billions`). Korištenjem modela linearne regresije provjerit ćemo koji regresori najviše utječu na izlaznu varijablu.

Pretpostavke modela:

- linearnost veze X i Y
- pogreške nezavisne, homogene i normalno distribuirane s $\epsilon \sim N(0, \sigma^2)$

Za predobradu podataka radimo sljedeće stvari:

- Izbacujemo nepotrebne regresore:
  - name

  - company.name
  - rank
  - location.gdp: više od pola vrijednosti su 0 (netočan podatak)
  - location.country.code i location.citizenship: koristimo location.region koji je veće granulacije
  - wealth.how.from emerging, wealth.how.was founder, wealth.how.was political: konstantne varijable
  - company.sector: jer ima previše različitih vrijednosti, koje kad bi one hot encodali bi dali previše stupaca
  - wealth.type_inherited, već sadržan u `inherited`
- ignoriramo uzorke s netočnim podacima (kriva dob)
- povećavamo granulaciju varijable relationship (slične/iste vrijednosti svodimu na jednu)
- izbacujemo manji broj uzoraka koji sadrži null vrijednosti

Sve kategorijske varijable obrađujemo tako da ih pretvorimo u dummy varijable. Svaka kategorijska varijabla predstavljena je svojom novom vlastitom varijablom koja poprima vrijednost 1 u slučaju da originalna kategorijska varijabla odgovara novoj varijabli, inače je vrijednost 0.

Za filtrirani podatkovni skup sve iznose `wealth.worth in billions` množimo s 1 + (kupovna moć dolara svedena na godinu 2014).

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

bill_data_clean <- bill_data %>% select(-one_of(exclude_cols)) %>% arrange(year)

bill_data_clean[["company.relationship"]] <- tolower(bill_data_clean[["company.relationship"]] )

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)

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

bill_data_clean <- bill_data_clean %>% drop_na()

bill_categorical <- bill_data_clean %>% select(where(is_character))
bill_numeric <- bill_data_clean %>% select(where(is.numeric))
bill_categorical_onehot = dummy_cols(bill_categorical, remove_first_dummy = TRUE, remove_selected_colum
bill_categorical_onehot <- bill_categorical_onehot[, colSums(bill_categorical_onehot) > 5]
bill_data_clean <- bind_cols(bill_numeric, bill_categorical_onehot)
```

Bitna pretpostavka multivarijatne linearne regresije je da ne postoji snažna linearna korelacija regresora modela. U ovom koraku provjerit ćemo postoje li parovi takvih regresora i otkloniti ih ako postoje. Odbacit ćemo sve regresore za koje postoji neki drugi regresor čiji je apsolutna vrijednost Pearsonovog koeficijenta korelacije veća od 0.9.

```r
correlation_threshold = 0.9
tmp <- corr_table <- 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))]

bill_data_clean <- remove_outliers(bill_data_clean, bill_data_clean$`wealth.worth in billions`)
wealth <- bill_data_clean$`wealth.worth in billions`
```

Prije stvaranja linearnog modela pogledajmo kojih 5 varijabli najviše linearno korelira sa `wealth.worth in billions`. Rezultate koje dobijemo ne možemo direktno koristiti za statističko zaključivanje, ali možemo kasnije usporediti linearne korelacije s rezultatima i zaključcima koje ćemo dobiti nakon stvaranja linearnog modela.

```r
w <- corr_table[, "wealth.worth in billions"]
w <- abs(w)

corr_wealth_vars <- w[order(w, decreasing = TRUE)]
cat("")
corr_wealth_vars[2:6]
```

```
## wealth.how.industry_Technology-Computer                      demographics.age
##                               0.08819941                            0.07454660
##           location.region_North America             company.relationship_owner
##                               0.06753705                            0.05941804
##           company.relationship_chairman
##                               0.05544800
```

```r
# 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", "wealth.type_inh
x <- bill_data_clean %>% select(-one_of(exclude_cols))
x[, c("company.founded", "demographics.age")] <- apply(x[, c("company.founded", "demographics.age")] , 1
x <- x[,order(colnames(x))]
```

Prvi linearni model stvorili smo naivno tako da smo iskoristili sve moguće regresore. Provjerom vrijednosti `adj.r.squared` saznat ćemo koliki postotak varijance u podacima opisuje stvoreni linearni model. Također, provjerit ćemo koji regresori objašnjavaju najveći postotak varijance tako da ih poredamo po p vrijednostima.

```r
cat("Ukupan broj regresora:", length(colnames(x)), "\n")
```

```
## Ukupan broj regresora: 47
```

```r
p_value_column <- 4
model_all_vars <- lm(wealth ~ . , x)
sa <- summary(model_all_vars)
cat("Postotak varijance objašnjen linearnim modelom", sa$adj.r.squared * 100, "%\n")
```

```
## Postotak varijance objašnjen linearnim modelom 7.799249 %
```

```
coef <- sa$coefficients
coef_sorted <- coef[order(coef[,p_value_column]),]
cat("Prvih 5 regresora sortiranih uzlazno po p vrijednosti:\n")
```

```
## Prvih 5 regresora sortiranih uzlazno po p vrijednosti:
```

```
coef_sorted[1:5,]
```

```
##                               Estimate Std. Error   t value     Pr(>|t|)
## demographics.age             1.2566438  0.2373435  5.294621 1.327898e-07
## company.type_subsidiary     -2.5848662  0.7346133 -3.518676 4.437019e-04
## company.relationship_owner -0.7681034  0.2206993 -3.480317 5.118976e-04
## company.type_new            -1.3695585  0.4038813 -3.390993 7.103700e-04
## company.type_privatization -1.6074880  0.4857169 -3.309517 9.516929e-04
```

U ovom slučaju, najveći postotak varijance u podacima za izlaznu varijablu `wealth.worth in billions` objašnjava regresor `demographics.age`. Trenutan model objašnjava svega 7.7% varijance u podacima (Adjusted $R = 0.07766$) za reakciju `wealth.worth in billions`. Na žalost, ovaj model ne objašnjava veliki dio varijance u podacima.

Sada ćemo pokušati pronaći najbolje prediktore na sljedeći način: stvarat ćemo model linearne regresije za svaki regresor pojedinačno, i očitavati koliko oni statistički značajno objašnjavaju varijancu u podacima (očitavamo p vrijednosti svakog modela nakon fittanja).

```
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]]) # napravi lienarni model s jednim regresorom
  summary_model = summary(model)

  filtered_col_names <- append(filtered_col_names, col_name)
  r_squares <- append(r_squares, summary_model$r.squared)
  ps <- append(ps,  summary_model$coefficients[,4][2])
}

df_g_squares=data.frame(filtered_col_names, r_squares, ps)
df_top_predictors = df_g_squares[order(df_g_squares$ps), ]
df_top_predictors[1:n, ]
```

```
##                    filtered_col_names   r_squares           ps
## 43 wealth.how.inherited_not inherited 0.023570667 5.980060e-12
## 42        wealth.how.inherited_father 0.018810687 8.312478e-10
## 12                  demographics.age 0.012642216 5.040078e-07
## 6          company.relationship_owner 0.011850950 1.150257e-06
## 47      wealth.type_self-made finance 0.006974850 1.938178e-04
## 30          wealth.how.industry_Media 0.005597208 8.453764e-04
## 17      location.region_North America 0.004616901 2.442123e-03
## 15      location.region_Latin America 0.004530739 2.682592e-03
## 2       company.relationship_chairman 0.004348546 3.273294e-03
## 4        company.relationship_founder 0.004031686 4.633888e-03
```

```
top_n_predictors_one_var_lin <- df_top_predictors[1:n, "filtered_col_names"]
```

Možemo zaključiti da kad bismo morali napraviti linearni model koji najbolje predviđa reaktor `wealth.worth in billions`, odabrali bismo upravo regresor `wealth.how.inherited_not inherited`. Međutim ako želimo napraviti multivarijatni linearni model, nije nužno istina da će najbolji model biti onaj za koji uzmemo prvih n regresora iz trenutne liste. Problem koji se može pojaviti takvim pristupom odabira regresora je da postoje regresori koji su međusobno zavisni (iako smo već prethodno otklonili jako zavisne regresore).

Najbolje regresore također možemo pronaći ANOVA-om. Kada dodamo ili izbrišemo prediktivnu varijablu iz linearne regresije, želimo znati je li ta promjena poboljšala model ili nije. ANOVA uspoređuje dva regresijska modela i javlja jesu li značajno različiti. Spojit ćemo najbolje regresore koje smo dobili ANOVA-om i najbolje regresore dobivene u prethodnom koraku da stvorimo novi linearni model s manje regresora.

```
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]

top_predictors = c(top_predictors_anova, top_n_predictors_one_var_lin)
top_predictors <- top_predictors[!duplicated(top_predictors)]

cat("Broj regersora", length(top_predictors))
```

```
## Broj regersora 14
```

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

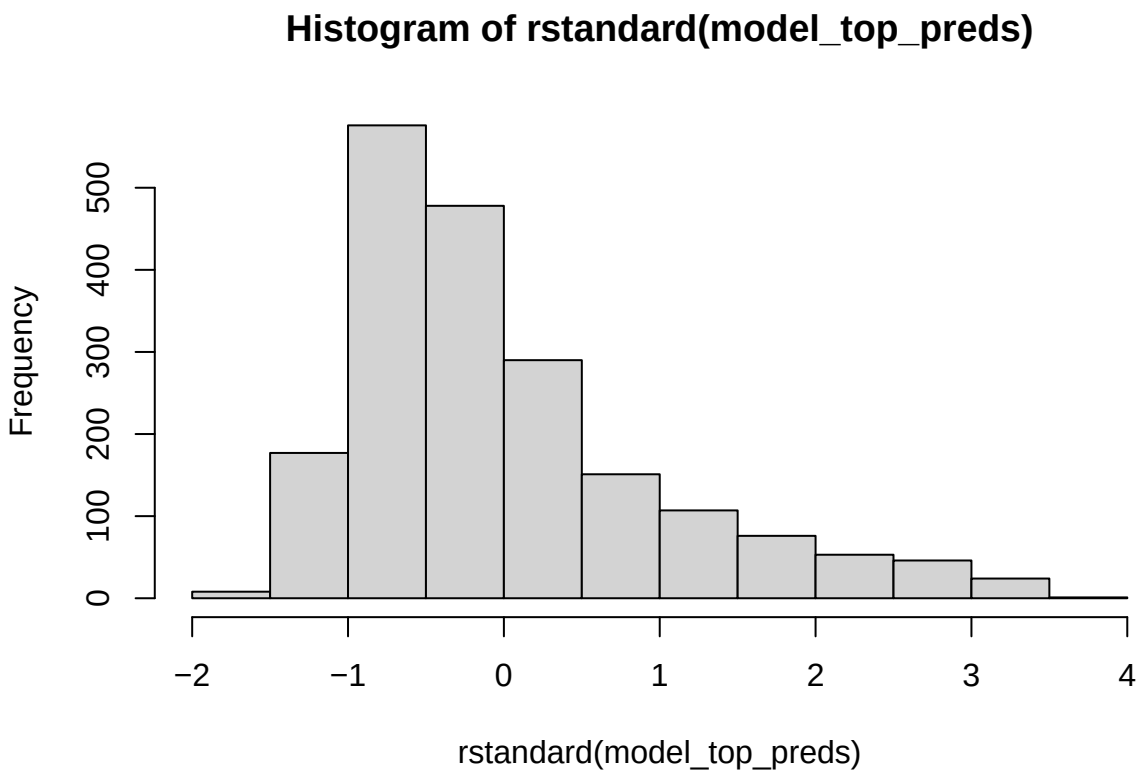
```
##
## Call:
## lm(formula = wealth ~ ., data = x[, top_predictors])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6846 -1.0576 -0.3915  0.6097  5.2619
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             2.12286    0.16430  12.921  < 2e-16
## company.relationship_owner             -0.65155    0.20385  -3.196 0.001415
## demographics.age                        1.22799    0.23058   5.326 1.12e-07
## company.relationship_founder           -0.23270    0.15168  -1.534 0.125164
## wealth.how.inherited_father             0.20062    0.11970   1.676 0.093894
## company.type_subsidiary                -1.21610    0.61610  -1.974 0.048538
## `location.region_Latin America`        -0.49507    0.14101  -3.511 0.000456
## company.relationship_chairman          -0.29599    0.18397  -1.609 0.107798
## `wealth.type_privatized and resources`  0.41470    0.18492   2.243 0.025032
## `wealth.how.industry_Technology-Computer` 0.23578  0.13622   1.731 0.083624
## `wealth.type_founder non-finance`       0.46838    0.18624   2.515 0.011983
## `wealth.how.inherited_not inherited`   -0.55339    0.16907  -3.273 0.001082
## `wealth.type_self-made finance`         0.29277    0.17512   1.672 0.094717
## wealth.how.industry_Media               0.46151    0.12976   3.557 0.000385
## `location.region_North America`         0.07785    0.07251   1.074 0.283157
```

```
##
## (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_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 1972 degrees of freedom
## Multiple R-squared:  0.06803,    Adjusted R-squared:  0.06141
## F-statistic: 10.28 on 14 and 1972 DF,  p-value: < 2.2e-16
```

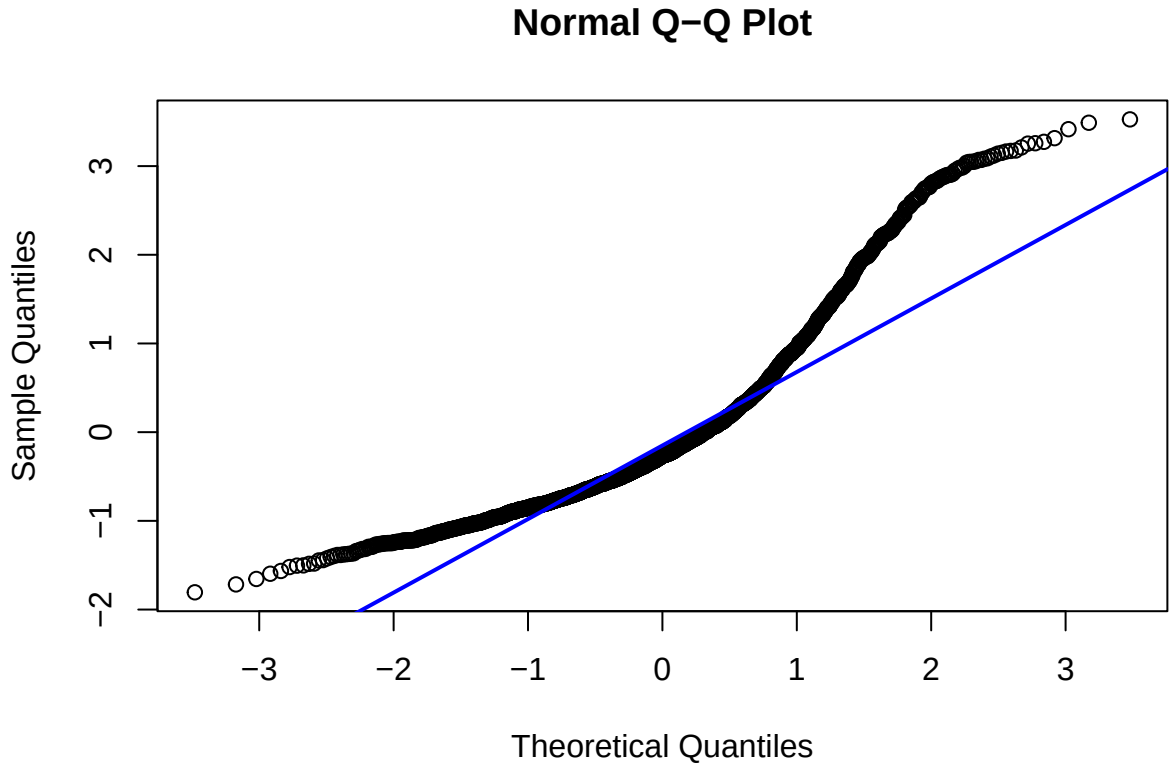
Pokazali smo da smanjenjem broja regresora na 14 i dalje objašnjavamo usporedivo veliki dio varijance (6.1%, originalno 7.7%). Ovisno o namjeni i potrebama možemo se opredijeliti za složeniji ili jednostavniji model. Jednostavniji model je preferiraniji ako je relativno dobar kao neki alternativni složeniji model.

Za kraj, provjerit ćemo pretpostavku linearnog modela (normalnost reziduala) grafički i korištenjem Kolmogorov-Smirnov i Lilliefors testa.

```
require(nortest)
hist(rstandard(model_top_preds))
```




```
qqnorm(rstandard(model_top_preds))
qqline(rstandard(model_top_preds), col = "blue", lwd = 2)
```




```
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 = 0.12694, 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 = 0.12695, p-value < 2.2e-16
```

Iz histogram se može naslutiti da se distribucija reziduala ne ravna po normalnoj distribuciji. Vrijednosti nisu centrirane oko nule i uočavamo debele desne repove. Iz qq grafa jasno uočavamo problem teškog desnog repa i manje problematičnog laganog lijevog repa. Ovaj graf dodatno potvrđuje da se reziduali ne ravnaju po normalnoj distribuciji. Konačno, oba testa za normalnost ukazuju da se reziduali ne ravnaju po normalnoj distribuciji jer je p vrijednost je manja od 0.05.

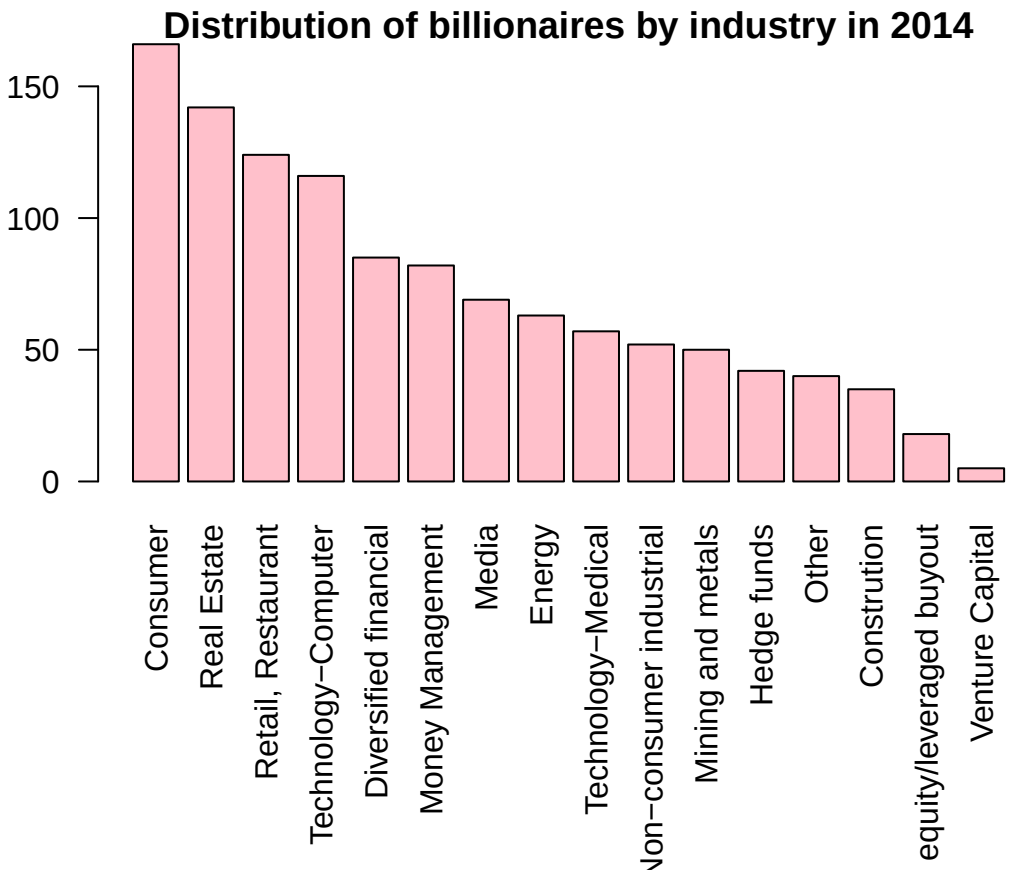
S obzirom na to da je pretpostavka linearnog modela prekršena i da linearan model objašnjava svega 6.2% varijance za reaktor `wealth.worth in billions` odbacujemo mogućnost da linearnim modelom previđamo bogatstvo koristeći preostale varijable iz skupa podataka.

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, jer su vremenski najrelevantniji. Uz taj skup milijardera, uzet ćemo i skup milijardera iz 2014. koji nisu bili na popisu 2001., odnosno novonastale milijardere. Na tom skupu vidjet će se u kojim industrijama je nastalo najviše milijardera. Za kraj ćemo uzeti u razmatranje i skup milijardera koji su bili na popisu 2001. godine, ali zbog određenog razloga više nisu. Tu ćemo vidjeti koje su industrije u tom razdoblju izgubile najviše milijardera.

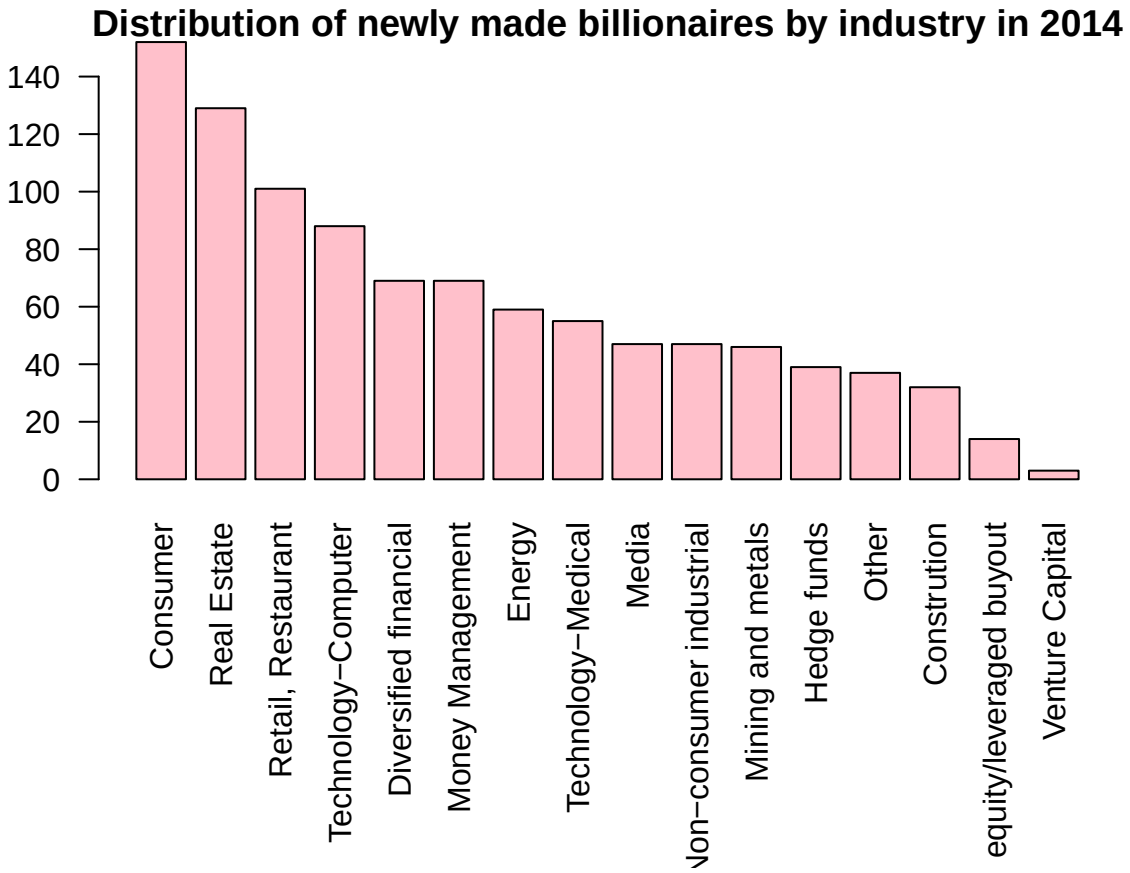
```
non_inherited_2014 <- non_inherited[non_inherited$year == 2014,]
non_inherited_2001 <- non_inherited[non_inherited$year == 2001,]
non_inherited_2014_new = bill_data[FALSE,]
non_inherited_2001_old = bill_data[FALSE,]

# 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(10,7,1,1))
barplot(sort(table(subset(non_inherited_2014$wealth.how.industry, non_inherited_2014$wealth.how.industry
        main = "Distribution of billionaires by industry in 2014",
        las = 2, col="pink")
```



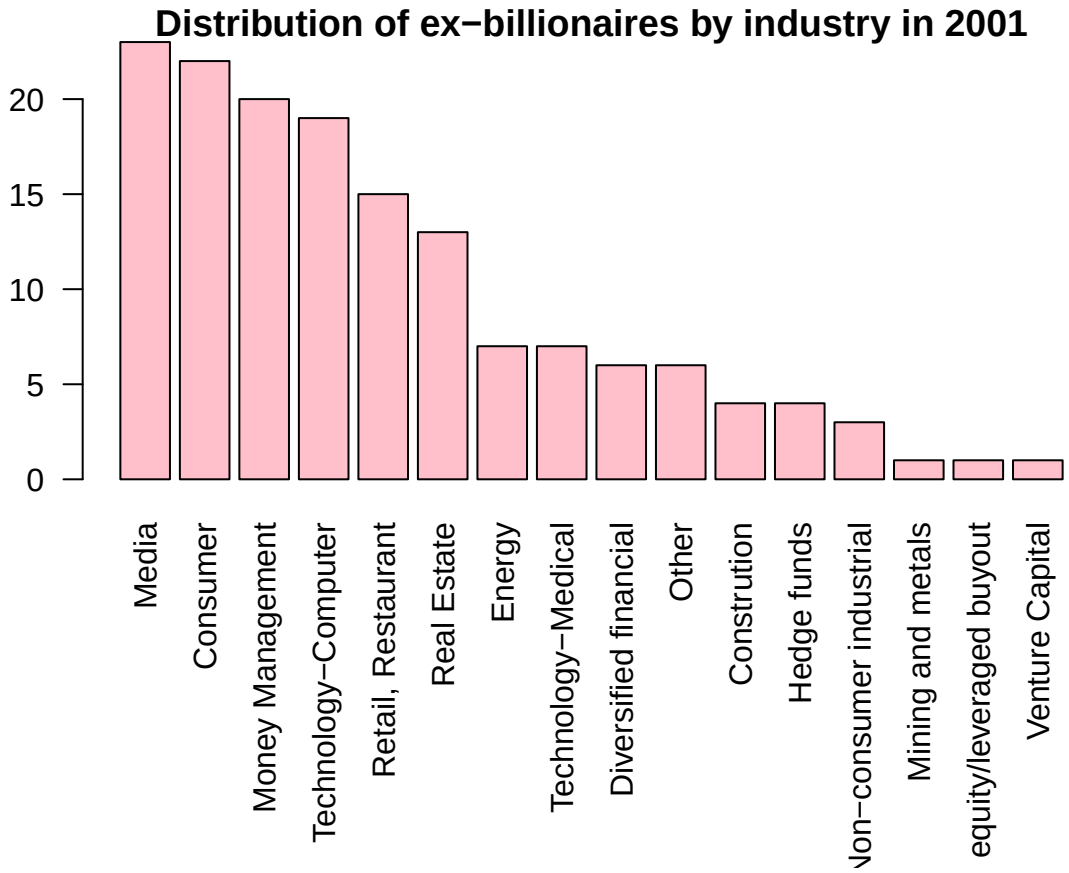
Iz stupčastog grafa je vidljivo da su tri najzastupljenije industrije maloprodaja (trgovački lanci, lanci restorana), trgovina nekretninama i računalna tehnologija.

```
par(mar=c(10,5,1,1))
barplot(sort(table(subset(non_inherited_2014_new$wealth.how.industry, non_inherited_2014_new$wealth.how
        main = "Distribution of newly made billionaires by industry in 2014",
        las = 2, col ="pink")
```



Ako usporedimo ovaj graf s prethodnim može se vidjeti da su vrlo slični, jedina razlika je u poretku medijske industrije. To nam govori da se broj milijardera po industrijama mijenja otprilike istom brzinom, odnosno da industrije s najviše milijardera dobivaju najveći broj novih milijardera (i obrnuto).

```
par(mar=c(10,5,1,1))
barplot(sort(table(subset(non_inherited_2001_old$wealth.how.industry, non_inherited_2001_old$wealth.how
        main = "Distribution of ex-billionaires by industry in 2001",
        las = 2, col="pink")
```



Na posljednjem grafu možemo potvrditi prethodno uočeno kretanje medijske industrije. Za razliku od ostalih industrija, broj milijardera u medijskoj industriji jedini je toliko značajno pao da je medijska industrija pala u ukupnom poretku. Sukladno tome na ovom grafu vidimo da je medijska industrija doživjela neproporcionalan pad broja milijardera. Međutim, medijska industrija nije među najvećim industrijama tako da ne utječe na zaključak, odnosno tri najzastupljenije industrije nisu se promijenile.

Zaključno, industrije koje se mogu predložiti na temelju ovih grafova su ponajprije maloprodaja, trgovanje nekretninama i računalna tehnologija. Te industrije najbolji su odabir za početak karijere s ciljem postajanja milijarderom ponajprije zbog najveće količine milijardera u tim industrijama (i sukladno tome najvećeg broja novonastalih milijardera).

5. Jesu li muškarci milijarderi statistički značajno bogatiji od žena milijardera?

S obzirom na to da muškaraca milijardera ima značajno više nego žena (rezultate smo dobili u deskriptivnoj analizi), čak 10 puta više, zanima nas jesu li onda oni i uspješniji,odnosno bogatiji u prosjeku od žena milijardera.

Postoje nedostajuće vrijednosti koje moramo izbaciti u varijabli demographics.gender(NA).

```
bill_data_gender= bill_data_copy%>% filter(!is.na(demographics.gender))



women = bill_data_gender[bill_data_gender$demographics.gender == "female",]
men = bill_data_gender[bill_data_gender$demographics.gender == "male",]
```

```
women_mean = mean(women$wealth.worth.in.billions)
women_mean
```

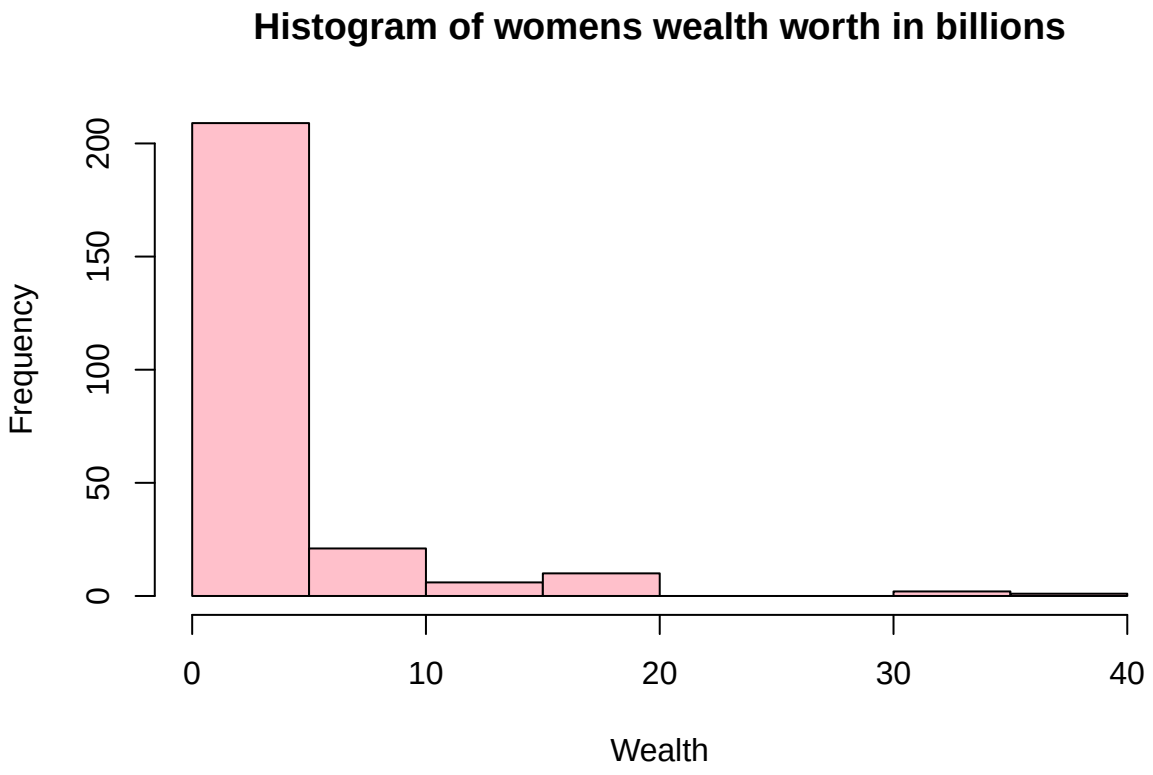
```
## [1] 3.819277
```

```
men_mean = mean(men$wealth.worth.in.billions)
men_mean
```

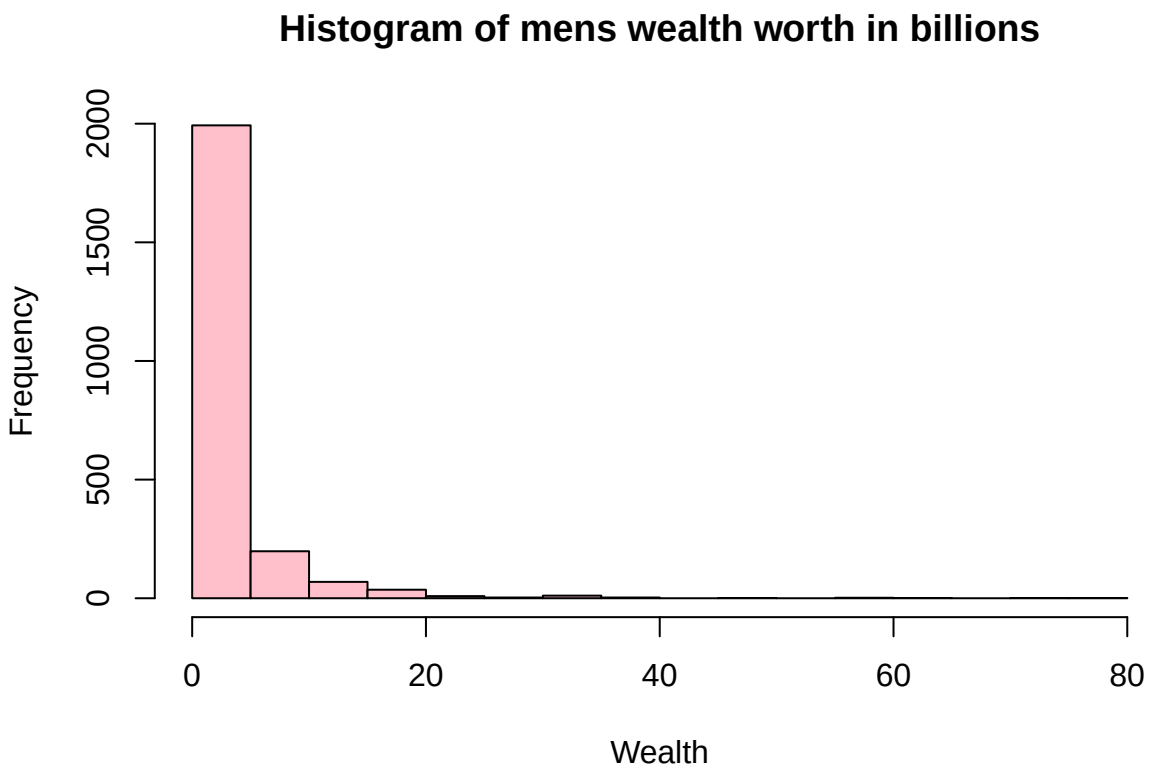
```
## [1] 3.516881
```

Izračunali smo srednje vrijednosti bogatstva žena i muškaraca. Iz ovih rezultata možemo vidjeti da se rezultati ne razlikuju puno, no treba provjeriti je li ova mala razlika statistički značajna. Kako bi mogli provesti t-test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Obzirom na to da razmatramo dva uzoraka različitih spolova, možemo pretpostaviti njihovu nezavisnost. Sljedeći korak je provjeriti normalnost podataka koju najčešće provjeravamo: histogramom te KS-testom (kojim provjeravamo pripadnost podataka distribuciji).

```
hist(women$wealth.worth.in.billions,
     main='Histogram of womens wealth worth in billions',
     xlab='Wealth',
     col="pink")
```




```
hist(men$wealth.worth.in.billions,
     main='Histogram of mens wealth worth in billions',
     xlab='Wealth',
     col="pink")
```



Iz histograma jasno vidimo da se podaci ne ravnaju po normalnoj distribuciji. Ipak, uočeno je potrebno dodatno ispitati koristeći Kolmogorov–Smirnov test kojim se utvrđuje ravna li se distribucija po normalnoj razdiobi.

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

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

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

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  men$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.

Potrebno je koristiti neparametarski test Mann–Whitney U test, koji se koristi kada se podaci se ravnaju po istim distribucijama i uzorci su nezavisni iz jedne i druge populacije (u našem datasetu osoba ima dodijeljen spol male ili female).

Hipoteze glase:

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

```r
wilcox.test(women_mean, men_mean, alt = "greater")
```

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

S obzirom na to da je dobivena p-vrijednost > 0.05, nemamo temelja za odbaciti nultu hipotezu, odnosno muškarci i žene se statistički značajno ne razlikuju u količini bogatstva.