```r
#### OPŠTE INFORMACIJE O BAZI ####


# Baza podataka je preuzeta sa sledećeg linka:
# https://github.com/fivethirtyeight/data/tree/master/alcohol-consumption

# U datoj bazi možemo pronaći podatke o potrošnji alkohola u 193 zemlje tokom 2010. godine.
# Potrošnja alkohola je iskazana kroz četiri varijable:

# 1. prosečna potrošnja piva po glavi stanovnika,
# koja je izražena kroz broj konzumiranih limenki piva (cans of beer),
# u ovoj anlizi ova varijabla će nositi naziv potrošnja limenki piva;

# 2. prosečna potrošnja vina po glavi stanovnika,
# koja je izražena kroz broj konzumiranih čaša vina(glasses of wine),
# u ovoj anlizi ova varijabla će nositi naziv potrošnja čaša vina;

# 3. prosečna potrošnja žestokih pića po glavi stanovnika
# koja je izražena kroz broj konzumiranih čašica žestokih pića (shots of spirits),
# u ovoj anlizi ova varijabla će nositi naziv potrošnja čašicia žestokih pića:

# 4. prosečan unos čistog alkohola po glavi stanovnika, koji je izražen u litrima,
# i u ovoj analizi ova varijabla će nositi naziv unos čistog alkohola.

# Važno je napomenuti da su čaše, čašice i limenke samo standardizovane mere
# za uobičajeni način konzumiranja datih pića a ne pravi podaci o načinu njihovog konzumiranja.
# Primera radi ako je neko popio dve litre piva to je iskazano preko 4 limenke,
# bez obzira na to da li je to pivo u realnosti konzumirano putem krigle, flaše, ili limenke.


####  UČITAVANJE I UPOZVAVANJE SA KARATERISTIKAMA BAZE ####


# originalna baza
orig_baza <- read.csv("data/drinks.csv", stringsAsFactors = FALSE)

# baza za sređivanje
baza_n <- orig_baza

# Upoznavanje sa osnovnim karateristikama baze
ncol(baza_n)
nrow(baza_n)

tail(baza_n, 10)
head(baza_n, 10)

summary(baza_n)
str(baza_n)

#####  SREĐIVANJE BAZE ####

# 1. Provera prisutnosti nedostajućih vrednosti.

is.na(baza_n) # klasičan način
sum(is.na(baza_n)) # pregledniji način, ako je vrednost nula onda ih nema

# Očigledno da nema nedostajućih vrednosti u NA formatu.
# Ali to ne znači da one stvarno ne postoje.

# Sve opservacije koje imaju vrednost nula,
# su potencijalno nedostajuće vrednosti ili rezultat lošeg merenja.
# Primera radi teško je zamisliti zemlju u kojoj stvarno nema nikakve potrošnje alkohola,
# i gde nijedan njen stanovnik ne konzumira ni kap piva, vina ili žestokih pića.
# Čak i kad bi postojala zabrana konzumiranja pića,
# to ne znači da bi se ona nužno poštovala u realnim okolnostima.
# Zbog toga smatram je potrebno eliminisati sve observacije tog tipa.
```

```r


# 2. Proces eliminacije

baza1 <- replace(baza_n, baza_n == 0, NA)
baza <- na.omit(baza1)
summary(baza)

# Prvo su zamenjene sve 0 vrednosti sa NA vrednostima.
# Nakon toga je primenjena funkcija na.omit,
# koja eliminiše sve observacije koje imaju vrednos NA.
# Na ovaj način je eliminisatno 38 zemalja,
# sa potencijalno "problematičim vrednostima".
# Broj elimnisanih zemalja može delovati kao preobiman,
# ali ne treba zaboraviti da da je,
# ova baza obuhvatila skoro sve zemlje sveta i da samim tim
# imamo idalje prilično dobar uzorak,
# u kome zasigurno nema nedostajućih vrednosti.


#### ZANIMLJIVOSTI ####


# U ovom delu su istaktnuti određeni zanimlijvi podaci,
# koji nisu rezultat ozbiljne statističke analize,
# već prostog izvlačenja podataka iz baze.
# Pa samim tim nemaju status istraživačkog nalaza,
# već zanimlijve, a možda i korisne infomracije.

# Za sve podatke koji će u ovom delu biti istaknuti,
# se podrazumeva da je potrošnja računata po glavi stanovnika.

# 1. Gde je najviše konzumiran alkohol tokom 2010. godine,
# U Srbiji, Bugarskoj ili Rusiji?

baza[baza$country == "Serbia", ]
baza[baza$country == "Bulgaria", ]
baza[baza$country == "Russian Federation", ]

# Očigledno je najviše konzumiran u Rusiji (11.5 litara čistog alkohola),
# u kojoj dominira potrošnja čašica žestokih pića (326).
# Mada je zanimljiv podatak da je u Srbiji
# konzumirano znatno više limenki piva (283) i čaša vina (127),
# u odnosu na Rusiju (247/73) i Bugarsku (231/94).


# 2. U kojoj zemlji je najviše konzumirano pivo tokom 2010. godine?
baza[(which.max(baza$beer_servings)), c(1,2)]
# Odgovor je pomalo začuđujući, reč je o Namibiji.

# 3. U kojoj zemlji je najviše konzumirano vino tokom 2010. godine?
baza[(which.max(baza$wine_servings)), c(1,4)]
# Odgovor ne bi trebalo da nas čudi, reč je o Francuskoj koja je poznata po vinima.

# 4. U kojoj zemlji su najviše konzumirana žestoka pića tokom 2010. godine?
baza[(which.max(baza$spirit_servings)), c(1,3)]
# Odgovor isto može da bude začućujući, pošto je reč o Grenadi.

# 5. U kojoj zemlji je najviše konzumiran alkohol tokom 2010. godine.
baza[(which.max(baza$total_litres_of_pure_alcohol)), c(1,5)]
# Reč je o Belorusiji.Odgovor je verovatno očekivan.

# 6. Lista zemalja sa najmanje konzumiranim pivom tokom 2010. godine.
baza[baza$beer_servings == 1, c(1,2)]

# 7. Lista zemalja sa najmanje konzumiranim vinom tokom 2010.godine.
baza[baza$wine_servings == 1, c(1,4)]

# 8. Lista zemalja u kojoj su najmanje konzumirana žestoka pića tokom 2010. godine.
```

```r
baza[baza$spirit_servings == 1, c(1,3)]

# 9. Za kraj možemo videti u kojoj zemlji je najmanje konzumiran alkohol tokom 2010. godine?
baza[(which.min(baza$total_litres_of_pure_alcohol)), c(1,5)]
# Najmanje je konzumiran na Komorima.


#### ISTRAŽIVAČKA PITANJA ####

# 1. Da li postoji statistički značajna veza između:
# - potrošnje: limenki piva i čaša vina
# - potrošnje: čašica žestokih pića i limenki piva
# - potrošnje: čaša vina i čašica žestokih pića

# 2. Koliko dobro mere potrošnje: limenki piva, čaša vina i čašica žestokih pića,
#  predviđaju ukupan unos čistog alkohola.

# Za prvo istraživačko pitanje koristiće se korelacioni testovi,
# a za drugo metod višestruke regresije.


#### PROVERA NORMALNOSTI DISTRIBUCIJE ####


# Ali pre primene pomenutih statističkih tehnika,
# treba proveriti normalnost distribucija datih varijabli,
# radi odabira adekvatnih stastistničkih testova.

# To radimo pomoću Shapiro-Wilk testa.

shapiro.test(baza$beer_servings)
shapiro.test(baza$spirit_servings)
shapiro.test(baza$wine_servings)
shapiro.test(baza$total_litres_of_pure_alcohol)

# p vrednost za sve varijable je znanto manja od 0.01
# usled čega odbacujemo nultu hipotezu u korist alterativne,
# i dolazimo do zakjučka da date varijable nemaju normalnu distribuciju,
# pa je poželjno koristiti neparametraske tehnike tamo gde je to moguće.

## Histogrami ##

# Pomoću histograma možemo grafički
# predstaviti distibuciju datih varijabli,
# kao bismo imali stekli što bolji uvid o njima.

## Instalacija i pozivanje ggplot paketa ##

# Instalacija nije neophodna ako je paket već instaliran,
# zato je i data u vidu komentara a ne komande, u narednom redu:
# install.packages("ggplot2")

# Isti princip će važiti za svako instaliranje paketa u ovoj analizi.

# pozivanje paketa je obavezan korak
library(ggplot2)

# 1. Histogram potrošnje lmenki piva.

ggplot(baza, aes( x = baza$beer_servings))+
  geom_histogram(aes(y = ..density..), bins = 12, colour = "white", fill = "grey75") +
  geom_density(aes(y = ..density..), colour = "blueviolet") +
  ggtitle("Potrošnja piva u 2010. godini") +
  xlab("broj konzumiranih limenki piva")+
  ylab("gustina")

# Ovde možemo zapaziti da je distribucija zakrivljena ulevo,
# iz čega možemo izvesti zaključak,
```

```r
# da je u velikom broju zemalja potrošnja limenki piva niska.

# 2. Histogram potrošnje čaša vina.

ggplot(baza, aes(x = baza$wine_servings))+
  geom_histogram(aes(y = ..density..), bins = 12, colour = "white", fill = "grey75") +
  geom_density(aes(y = ..density..), colour = "darkred") +
  ggtitle("Potrošnja vina u 2010. godini") +
  xlab("broj konzumiranih čaša vina") +
  ylab("gustina")

# Sličan je slučaj kao i sa pivom,
# samo što je u ovom slučaju zakrivljenje još intezivnije.
# Iz čega možemo zaključiti da je potrošnja čaša vina
# u još većm broju zemalja niska nego što je to slučaj sa pivom.

# 3. Histogram potrošnje čašica žestokih pića.

ggplot(baza, aes(x = baza$spirit_servings)) +
  geom_histogram(aes(y = ..density..), bins=12, colour = "white", fill = "grey75") +
  geom_density(aes(y =..density..), colour = "dodgerblue1") +
  ggtitle("Potrošnja žestokih pića u 2010. godini") +
  xlab("broj konzumiranih čašica žestokih pića") +
  ylab("gustina")

# Isti slučaj kao i sa vinom i pivom,
# distribucija je zakrivljena ulevo
# što znači da u većini zemalja imamo,
# nisku potrošnu čašica žestokog pića.

# 4. Histogram unosa čistog alkohola.

ggplot(baza, aes(x = baza$total_litres_of_pure_alcohol)) +
  geom_histogram(aes(y = ..density..), bins = 12, colour = "white", fill = "grey75") +
  geom_density(aes(y = ..density..), colour = "slateblue") +
  ggtitle("Unos čistog alkohola u 2010. godini") +
  xlab("čist alkohol izražen u litrima") +
  ylab("gustina")

# Zakrivljenje je znatno manje izraženo nego u predhodnim varijablama.
# U ovom slučaju imamo velki broj,
# kako zemalja sa niskim vrednostima unosa čistog alkohola,
# tako i zemalja sa srednjim vrednostima (opseg 5-10),
# dok je najmanji broj zemalja sa izrazito visokoim vrednostima ove varijalbe.


##### KORELACIJA #####

pvs <- data.frame(baza$beer_servings, baza$spirit_servings, baza$wine_servings)

# Matrica za korelaciju koja je sastavljena od
# varijabli nad kojima želimo da primenimo korelacione testove.


# Instalacija i pozivanje paketa za korelaciju
# install.packages("Hmisc")
library("Hmisc")

## Korelacioni testovi nad matricom ##

rcorr(as.matrix(pvs), type = c("spearman"))

n <- rcorr(as.matrix(pvs), type = c("spearman"))
print(n$P, digits = 15) # tačniji prikaz p vrednosti

## Nalazi korelacionih testova ##

# 1. Veza između potrošnje limenki piva i potrošnje čaša vina,
# istražena je pomoću Spirmanovog ro koeficijenta korelacije,
```

```r
# izračunata je pozitivna korelacija srednje jačine između dve promenjive,
# r = 0.62, n = 155, p < 0,01
# a to znači da sa porastom potrošnje limenki piva,
# raste i potrošnja čaša vina.

# 2. Veza između potrošnje čašica žestokog pića i potrošnje limenki piva
# istražena je pomoću Spirmanovog ro koeficijenta korelacije,
# izračunata je pozitivna korelacija srednje jačine između dve promenjive,
# r = 0.50, n = 155, p < 0,01
# a to znači da sa porastom potrošnje limenki piva,
# raste i potrošnja čašica žestokih pića.

# 3. Veza između čaša vina i čašica žestokog pića
# istražena je pomoću Spirmanovog ro koeficijenta korelacije,
# izračunata je pozitivna korelacija srednje jačine između dve promenjive,
# r = 0.39, n = 155, p < 0,01
# a to znači da sa porastom potrošnje čaša vina,
# raste i potrošnja čašica žestokih pića.


## Vizualizacija korelacije u ggplot-u (dijagrami raspršenosti) ##

# Instalacija i pozivanje ggplot-a
# install.packages("ggplot2")
library(ggplot2)

# Prikaz dijagrama

# 1. Dijagram korelacije potrošnje čaša vina i limenki piva

ggplot(baza, aes(x = baza$beer_servings, y = baza$wine_servings)) +
  geom_point(size = 3, shape = 2, colour = "blue") +
  ggtitle("Korelacija potrošnje limenki piva i čaša vina") +
  xlab("potrošnja limenki piva") +
  ylab("potrošnja čaša vina")

# 2. Dijagram korelacije potrošnje čašica žestokih pića i limenki piva

ggplot(baza, aes(x = baza$beer_servings, y = baza$spirit_servings)) +
  geom_point(size = 3, shape = 1, colour = "red3")+
  ggtitle("Korelacija potrošnje limenki piva i čašica žestokih pića") +
  xlab("potrošnja limenki piva") +
  ylab("potrošnja čašica žestokih pića")

# 3. Dijagram korelacije potrošnje čaša vina i čašica žestokih pića

ggplot(baza, aes(x = baza$wine_servings, y = baza$spirit_servings)) +
  geom_point(size = 3, shape = 0, colour = "purple") +
  ggtitle("Korelacija potrošnje čaša vina i čašica žestokih pića") +
  xlab("potrošnja čaša vina") +
  ylab("potrošnja čašica žestokih pića")

## Vizualizacija u ggcorrplot-u (kvadrati i krugovi) ##

# Instaliranje i pozivanje datog paketa paketa
# install.packages("ggcorrplot")
library(ggcorrplot)

#Sređivanje imena kolona i matrica za korelaciju
names(pvs) = c("potrošnja krigli piva", "potrošnja čašica žestokih pića", "potrošnja
čaša vina")
viz <- cor(pvs, method = "spearman")

#Vizalizacija korelacije
ggcorrplot(viz, lab = TRUE) # Prvi način, kvadrat
ggcorrplot(viz, method = "circle") # drugi način krug


#### REGRESIJA ####
```

```r


# Korelacioni testovi predstavljaju
# prvi korak za izradu,regresionih modela


# Zato prvo treba napraviti korelacionu matricu
# a nakon toga sprovesti korelacione testove.

reg <- data.frame(baza$total_litres_of_pure_alcohol, baza$beer_servings,
baza$wine_servings, baza$spirit_servings)

# Instalacija i pozivanje paketa za korelaciju
# install.packages("Hmisc")
library("Hmisc")

## Korelacioni testovi ##

rcorr(as.matrix(reg), type = c("spearman"))
m <-rcorr(as.matrix(reg), type = c("spearman"))
print(m$P, digits = 5)


## Nalazi ##

# Spirmanov  ro koeficijent korelacije,
# je pokazao da postoji pozitivna korelacija između varijable
# ukupnog unosa unosa čistog alkohola i svih drugih varijabli potrošnje alkohola.
# Najjača je korelacija sa potrošnjom limenki piva (r = 0.85).
# Dok je sa potrošnjom čaša vina(r = 0.70)
# i čašica žestokog pića(r = 0.66) ona umerene jačine.
# Iz svega pomenutog možemo izvesti zaključak
# da sa porastom potrošnje limenki piva, čaša vina i čašica žestokih pića,
# raste i ukupan unos čistog alkohola.
# Ovakav nalaz je logičan i očekivan.
# Sve pomenute veze su statistički značajne p < 0.01

# Iz ovakvih nalaza možemo zaključiti da je najsvrsihsodnije
# napraviti tri regresiona modela: jedan sa svim varijablama,
# drugi sa pivom i vinom i treći samo sa pivom.

# Izrada regresionih modela:

# 1. Prvi model, sve tri varijable:

Model1 <- lm(baza$total_litres_of_pure_alcohol ~ baza$wine_servings +
baza$spirit_servings + baza$beer_servings )
summary(Model1)

# Model je odličan, pošto objašnjava 88% varjabiliteta  ukupnog unosa čistog alkohla.
# r^2 = 0.88, F = 385, p < 0.01 (za ceo model i sve koeficijente)

# 2. Drugi model, bez žestokih pića:

Model2 <- lm (baza$total_litres_of_pure_alcohol ~ baza$wine_servings +
baza$beer_servings)
summary(Model2)

# I ovaj model je prilično dobar pošto objašnjava 75% varjabiliteta
# ukupnog unosa čistog alkholoa, sa dve varijable
# r^2 = 0.75, F = 234.5, p < 0.01 (za ceo model i sve koeficijente)

# 2.1. Vizualizacija drugog modela

ggplot(Model2,aes(y=baza$total_litres_of_pure_alcohol,x=baza$beer_servings,
color=baza$wine_servings)) +
  geom_point(size = 4) +
  stat_smooth(method = "lm", se = FALSE, colour = "red", size = 1.3 )+
  ggtitle("Model broj 2") +
```

```r
  xlab("potrošnja limenki piva") +
  ylab("ukupan unos čistog alkohola") +
  labs(color = "potrošnja čaša vina")

# 3. Treći model, samo limenke piva:

Model3 <- lm (baza$total_litres_of_pure_alcohol ~  baza$beer_servings)
summary(Model3)

# Ovaj model je takođe veoma dobar pošto objašnjava 66 % varjabiliteta
# ukupnog unosa čistog alkohola sa samo jednom varijablom
# r^2 = 0.66, F = 304.9, p < 0.01 (za ceo model i koeficijente)

# 3.1. Vizualizacija trećeg modela

ggplot(Model3, aes(y = baza$total_litres_of_pure_alcohol, x = baza$beer_servings)) +
  geom_point(size = 3, color = "azure4" )+
  stat_smooth(method = "lm",se = FALSE) +
  ggtitle("Model broj 3") +
  xlab("potrošnja limenki piva") +
  ylab("ukupna unos čistog alkohola")

# Zaključak

# Sva tri modela su odlična ali je prvi ipak najbolji,
# jer objašnjava najveći procenat varijabiteta ukupnog unosa čistog alkohola.
# Ovaj model pokazuje da su
# varijable potrošnje čaša vina, limenki piva i čašica žestokog pića,
# odlčni prediktori vrednosti varijable ukupnog unosa čistog alkohola.
# Takav nalaz je očekivan pošto je reč o najpopularnijim vrstama pića.

#### ZAVRŠETAK ANALIZE ####

# Analizu sproveo: Novak Tešić
# GitHub profil: https://github.com/NovakTesic
```