Analiza glazbenih preferencija i njihovog utjecaja na mentalno zdravlje

Ena Dvojak, Patrik Blašković, Mislav Hlupić, Eugen Vucelić

2025-01-26

Uvod

Muzikoterapija je terapijski pristup koji koristi glazbu i njezine elemente za smanjenje stresa, poboljšanje raspoloženja i jačanje mentalnog zdravlja. U tom se procesu koriste različiti glazbeni žanrovi prema preferencijama pojedinca, a glazba služi kao katalizator pozitivnih emocionalnih reakcija.

Case study: *Muzikoterapija i slušačke navike*

U analizu je uključen skup podataka sa 736 odgovora na anketu koja istražuje slušačke navike, preferirane glazbene žanrove, samoprocijenjeno mentalno zdravlje i druge relevantne čimbenike (dob, odabir streaming servisa, učestalost slušanja određenih žanrova, instrumentalne i skladateljske vještine). Cilj ovog projekta je razumjeti obrasce i povezanosti u podacima te ih statistički interpretirati kako bi se dobio uvid u ulogu glazbe u poticanju pozitivnog mentalnog stanja.

```
library(tidyverse)
library(modeest)
library(dplyr)
library(nortest)
library(ggplot2)
library(corrplot)
library(car)
library(caret)
```

```
source("R/barplots.R")
source("R/normality_tests.R")
source("R/multiple_var_regression.R")
```

```
music_data <- read_csv("data/processed/dataset_reduced.csv", show_col_types = FALSE)

head(music_data)
```

```
## # A tibble: 6 x 31
##     Age Primary streaming serv~1 `Hours per day` `While working` Instrumentalist
##   <dbl> <chr>                             <dbl> <chr>           <chr>
## 1    18 Spotify                             3   Yes             Yes
## 2    63 Pandora                             1.5 Yes             No
## 3    18 Spotify                             4   No              No
## 4    61 YouTube Music                       2.5 Yes             No
```

```
## 5    18 Spotify                               4   Yes             No
## 6    18 Spotify                               5   Yes             Yes
## # i abbreviated name: 1: ‘Primary streaming service‘
## # i 26 more variables: Composer <chr>, ‘Fav genre‘ <chr>, Exploratory <chr>,
## #   ‘Foreign languages‘ <chr>, BPM <dbl>, ‘Frequency [Classical]‘ <chr>,
## #   ‘Frequency [Country]‘ <chr>, ‘Frequency [EDM]‘ <chr>,
## #   ‘Frequency [Folk]‘ <chr>, ‘Frequency [Gospel]‘ <chr>,
## #   ‘Frequency [Hip hop]‘ <chr>, ‘Frequency [Jazz]‘ <chr>,
## #   ‘Frequency [K pop]‘ <chr>, ‘Frequency [Latin]‘ <chr>, ...
```

Deskriptivna statistika

Mjere centralne tendencije

Mjere centralne tendencije za numeričke varijable `Age` i `Hours per day` grupirane po primarnom streaming servisu. Mod je izostavljen zato što nije dobra mjera centralne tendencije za ovaj dataset. Najveća frekvencija sati slušanja i godina ispitanika nije nužno najbolji pokazatelj sredine podataka, pogtovo za streaming servise s manje korisnika.

Koristiti ćemo 10% podrezanu aritmetičku sredinu jer iako su potencijalni ekstremni podaci značajni za analizu navika slušanja glazbe, moramo uzeti u obzir i neozbiljne ispune ankete koje mogu značajno utjecati na sredinu podataka.

```
music_data %>%
  group_by(`Primary streaming service`) %>%
  summarise(
    count = n(),
    mean_age = mean(Age, na.rm = TRUE,trim = 0.1),
    median_age = median(Age, na.rm = TRUE),
    mean_hours = mean(`Hours per day`, na.rm = TRUE, trim = 0.1),
    median_hours = median(`Hours per day`, na.rm = TRUE)
  ) %>%
  arrange(desc(count))
```

```
## # A tibble: 6 x 6
##   ‘Primary streaming service‘ count mean_age median_age mean_hours median_hours
##   <chr>                       <int>    <dbl>      <dbl>      <dbl>        <dbl>
## 1 Spotify                       458     21.3       20         3.30            3
## 2 YouTube Music                  94     25.3       22         2.74            2
## 3 None                           72     28.7       23.5       2.54            2
## 4 Apple Music                    51     21.7       20         3.01            2
## 5 Other                          50     28.3       25         2.83            3
## 6 Pandora                        11     51.6       60         2.06            2
```

Brisanje outliera

Prije računanja mjere rasipanja i vizualizacije box plotom, pokušati ćemo ukloniti outliere iz varijabli `Age` i `Hours per day` zdravim razumom. Naime anketa sadrži podatke o korisnicima koji su upisali nerealne godine ili sate slušanja glazbe. Uzimajući u obzir da jedan dan ima 24 sata i da većina ljudi ne sluša glazbu dok spava, realna maksimalna granica za `Hours per day` je 24 sata - 8 sati sna = 16 sati slušanja dnevno (ovime smo uklonili ukupno 3 data pointa). Za dob ispitanika ćemo uzeti u obzir da stariji generalno nisu skloni ispunjavanju anekta stoga ćemo heuristički staviti maksimalnu granicu za `Age` na 70 godina (ovime smo uklonili ukupno 7 data pointa).

```
music_data <- music_data %>%
  filter(`Hours per day` <= 16, Age <= 70)
```

Mjere rasipanja

Mjere rasipanja za varijable **Age** i **Hours per day** grupirane po primarnom streaming servisu. Pomoću standardne devijacije i ranga donosimo zaključke da **Spotify** ima najveći rang godina (ima i najviše korisnika među ispitanicima općenito), dok **Pandora** ima najveću standardnu devijaciju godina među svojim korisnicima (čemu pridonosi činjenica da je Pandora najmanje zastupljena među ispitanicima).

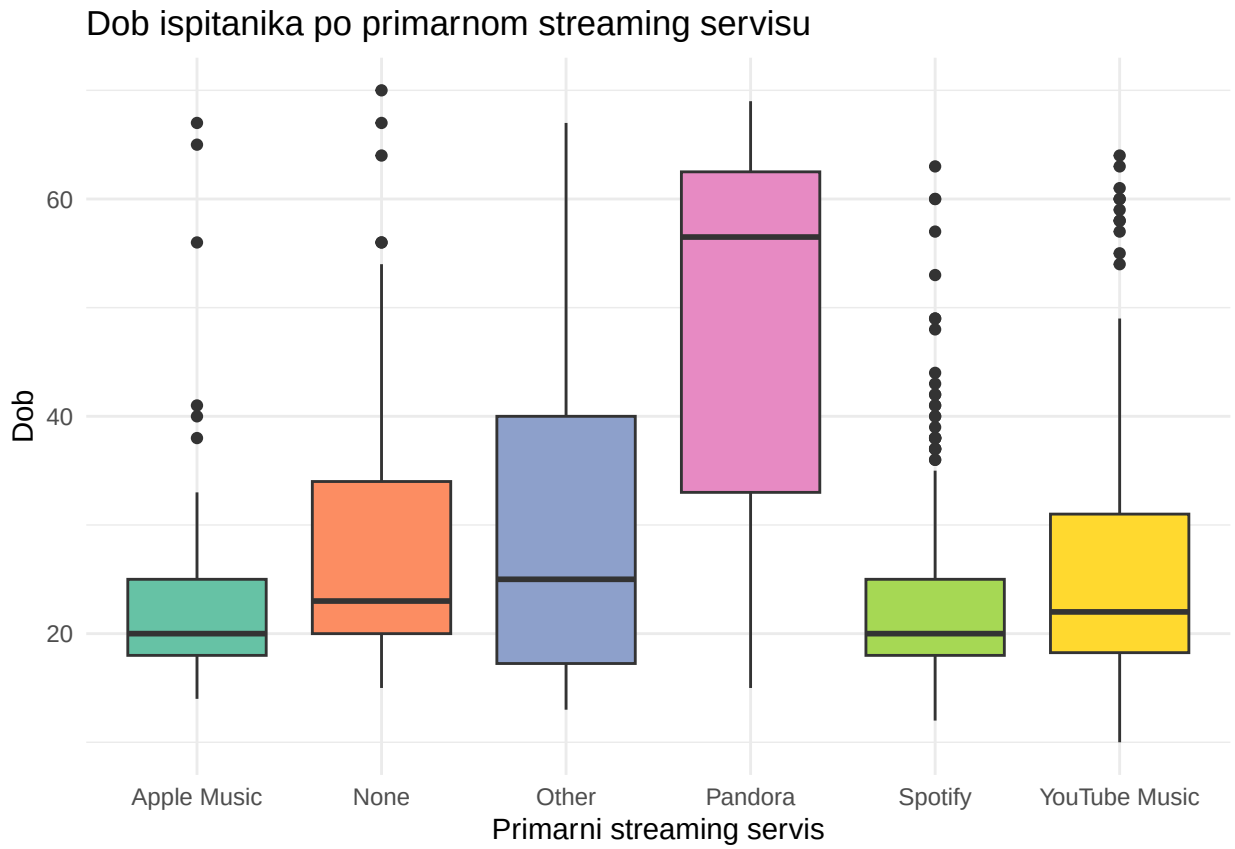
```
music_data %>%
  group_by(`Primary streaming service`) %>%
  summarise(
    sd_age = sd(Age, na.rm = TRUE),
    range_age = max(Age, na.rm = TRUE) - min(Age, na.rm = TRUE),
    sd_hours = sd(`Hours per day`, na.rm = TRUE),
    range_hours = max(`Hours per day`, na.rm = TRUE) - min(`Hours per day`, na.rm = TRUE)
  ) %>%
  arrange(desc(sd_age))
```

```
## # A tibble: 6 x 5
##   `Primary streaming service` sd_age range_age sd_hours range_hours
##   <chr>                         <dbl>     <dbl>    <dbl>       <dbl>
## 1 Pandora                       19.9         54     1.12         3
## 2 Other                         15.8         54     2.16        10
## 3 None                          13.8         55     2.79        12
## 4 YouTube Music                 13.7         54     2.74        14.8
## 5 Apple Music                   11.7         53     2.46        11.5
## 6 Spotify                        7.59        51     2.63        16
```

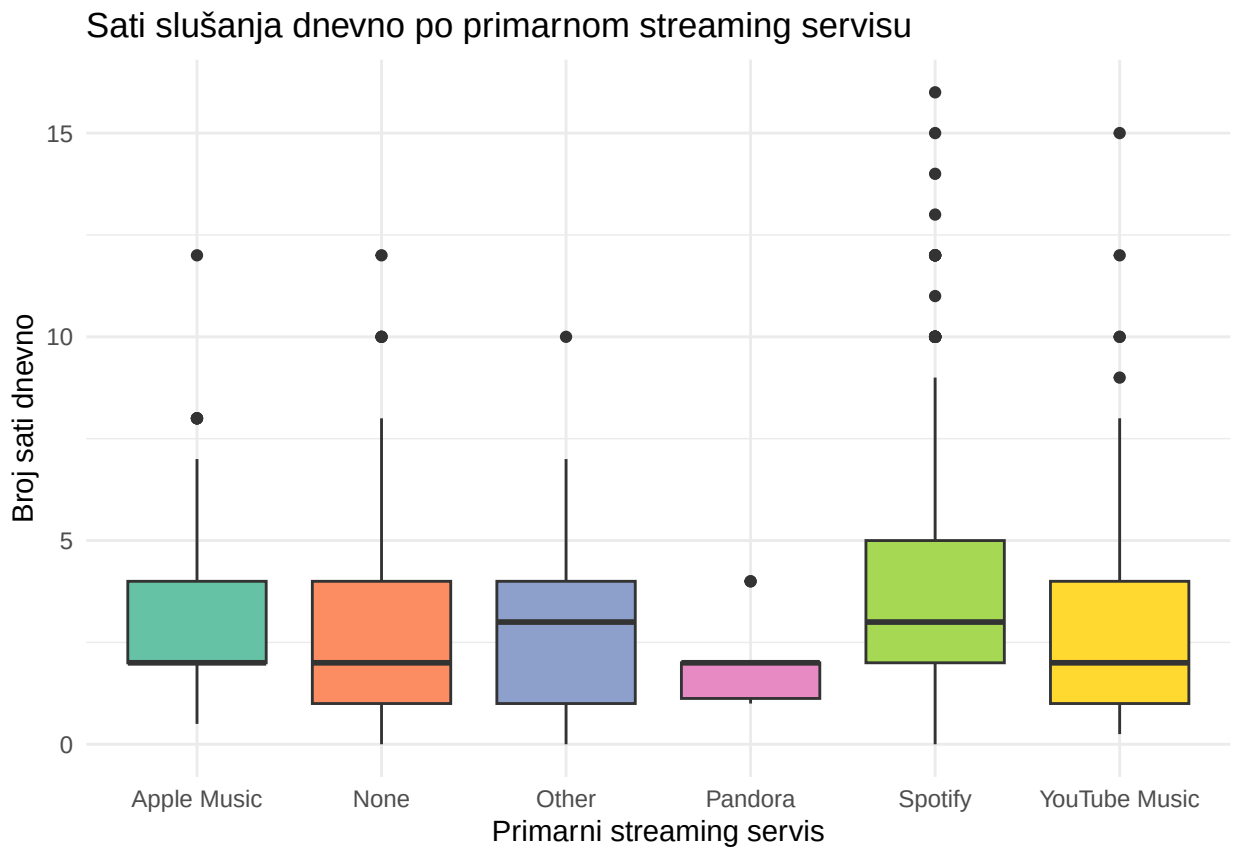
Vizualizacija boxplot

Iz box plotova vizualiziranih ispod možemo zaključiti da mlađi ispitanici preferiraju Spotify, Apple Music i YouTubeMusic dok stariji ispitanici preferiraju Pandoru ili alternativne servise. Spotify ima najaktivnije korisnike s najviše slušanih sati dnevno s najviše ljudi koji premašuju gornje whiskere boxplota. Outlieri prikazani boxplotom su zdravi i razumni podaci.

```
music_data %>%
  ggplot(aes(x = `Primary streaming service`, y = Age, fill=`Primary streaming service`)) +
  geom_boxplot() +
  labs(title = "Dob ispitanika po primarnom streaming servisu",
       x = "Primarni streaming servis",
       y = "Dob") + scale_fill_brewer(palette = "Set2") +
  theme_minimal() +
  theme(legend.position = "none")
```



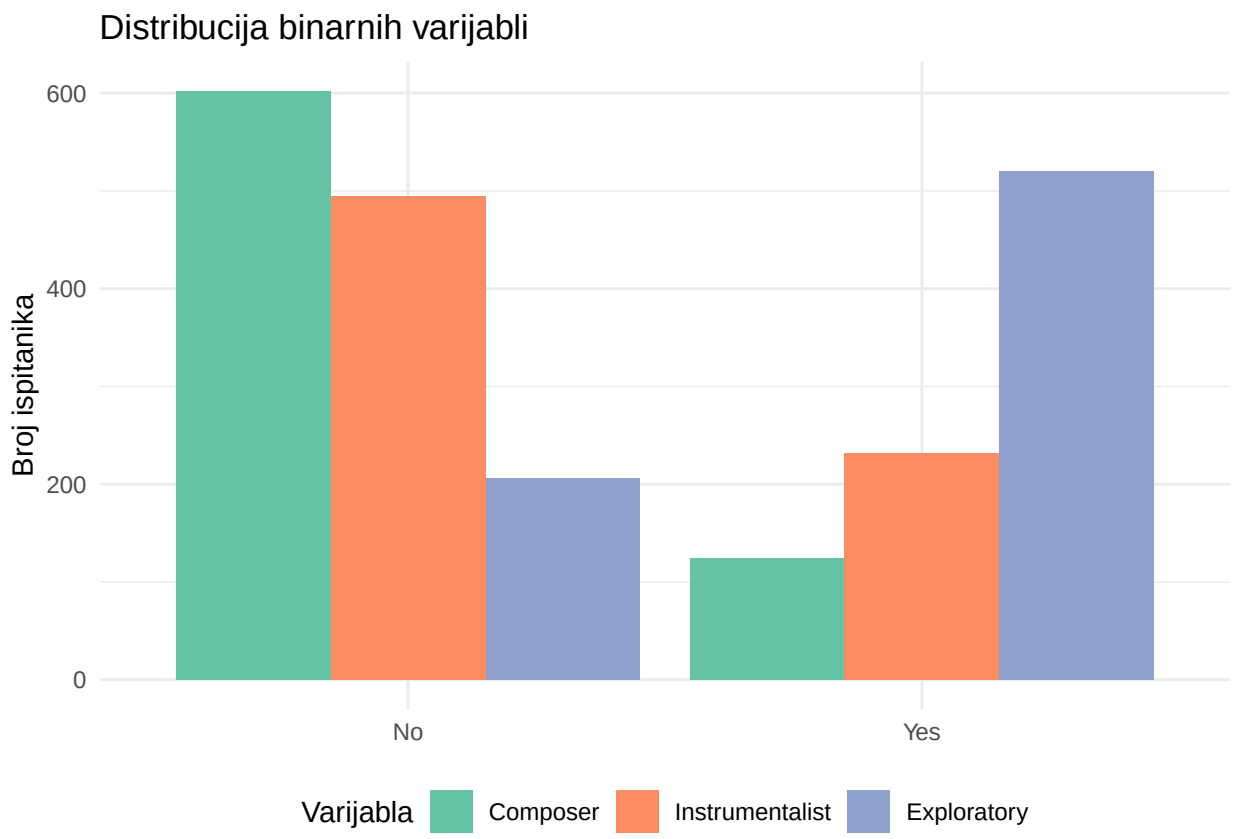

```
music_data %>%
  ggplot(aes(x = `Primary streaming service`, y = `Hours per day`, fill=`Primary streaming service`)) +
  geom_boxplot() +
  labs(title = "Sati slušanja dnevno po primarnom streaming servisu",
       x = "Primarni streaming servis",
       y = "Broj sati dnevno") + scale_fill_brewer(palette = "Set2") +
  theme_minimal() +
  theme(legend.position = "none")
```



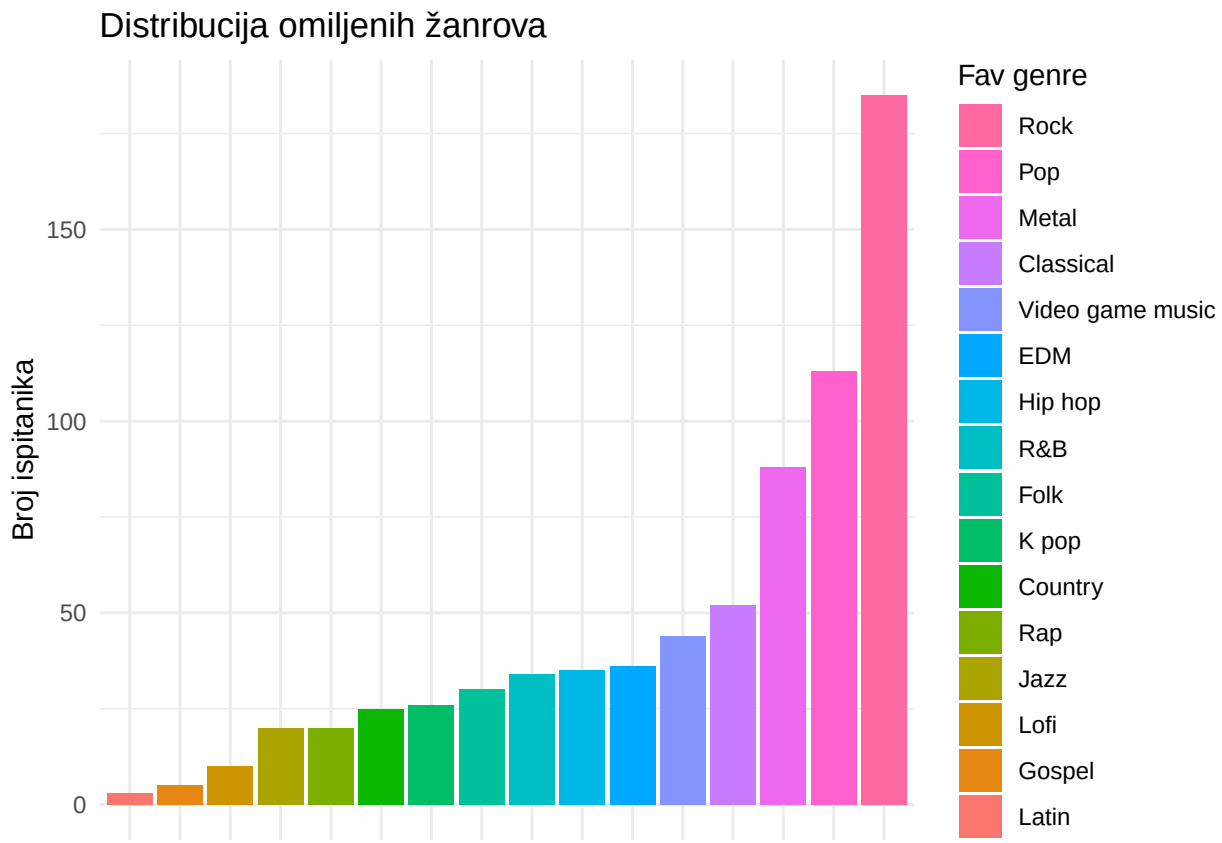
Vizualizacija barplot

Barplotom ćemo vizualiraiti neke kategoričke varijable koje će biti od značaja u daljnim analizama. Izabrane kategoričke varijable su `Instrumentalist`, `Composer`, `Fav genre` i `Exploratory`.

```
plot <- create_binary_plots(music_data)
print(plot)
```



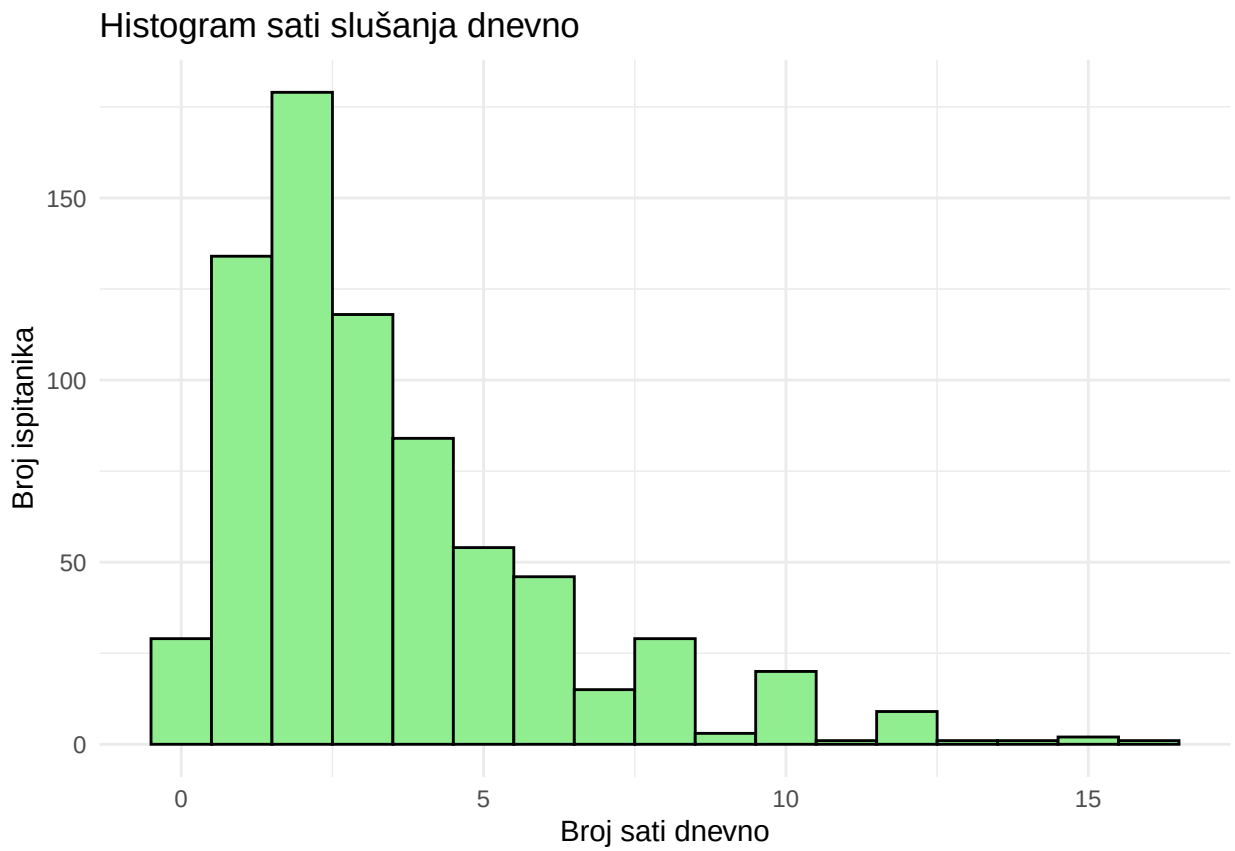

```
plot <- create_fav_genre_plot(music_data)
print(plot)
```



Histogrami i provjere normalnosti

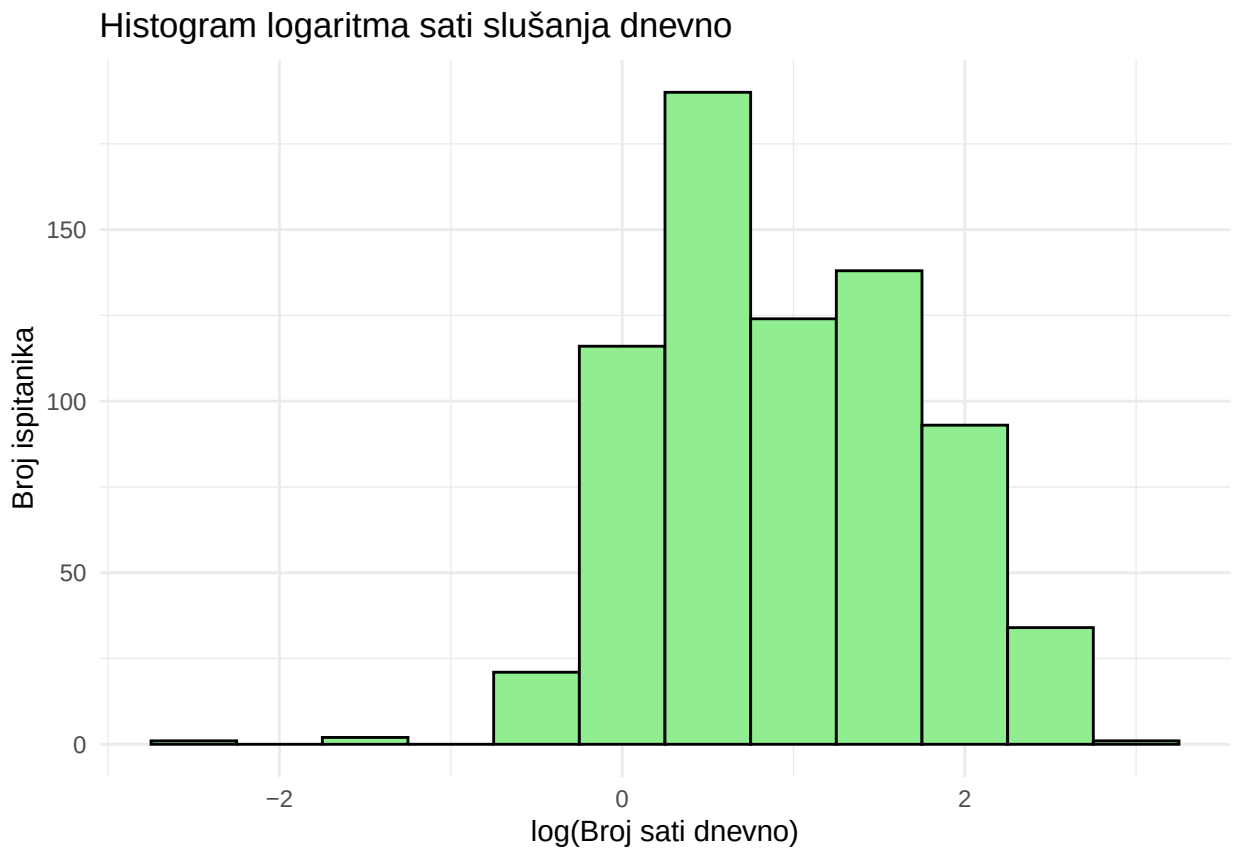
Za provjeru normalnosti distribucije numeričkih varijabli koristimo vizualne i statističke metode. Vizualno ćemo distribucije prikazati histogramima, dok ćemo statistički normalnost testirati **Lillieforsovom inačicom Kolmogorov-Smirnov testa**. Ovaj test je modifikacija standardnog Kolmogorov-Smirnov testa gdje se parametri normalne distribucije (srednja vrijednost i standardna devijacija) procjenjuju iz uzorka. Nulta hipoteza testa je da podaci dolaze iz normalne distribucije, a test provodimo na razini značajnosti $\alpha = 0.05$. Provjeru provodimo za numeričke varijable: `Hours per day` te samoprocjenjene skale `Anxiety`, `Depression`, `Insomnia`, `OCD`.

```
mean_hours <- mean(music_data$`Hours per day`, na.rm = TRUE)
sd_hours <- sd(music_data$`Hours per day`, na.rm = TRUE)

music_data %>%
  ggplot(aes(x = `Hours per day`)) +
  geom_histogram(binwidth = 1, fill = "lightgreen", color = "black") +
  labs(title = "Histogram sati slušanja dnevno",
       x = "Broj sati dnevno",
       y = "Broj ispitanika") +
  theme_minimal()
```



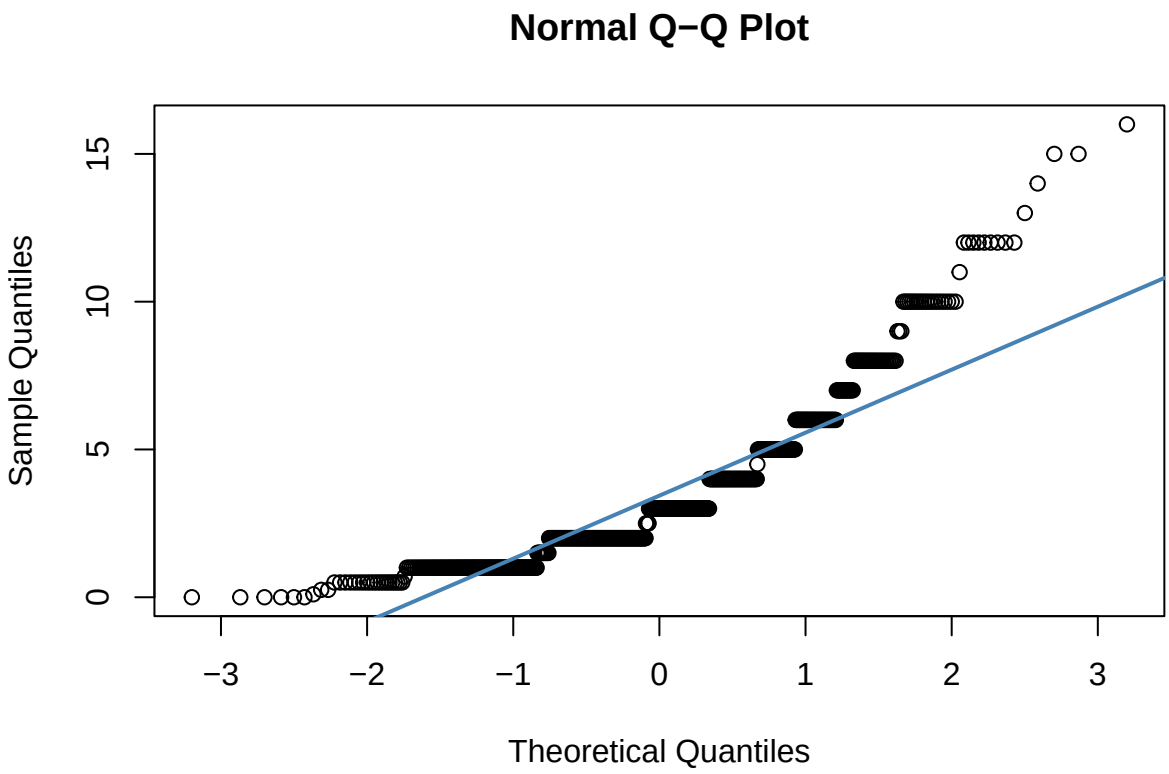
Provjerimo također jesu li podaci o dnevnom slušaju potencijalno log-normalni

```
music_data %>%
  filter(`Hours per day` > 0) %>%
  ggplot(aes(x = log(`Hours per day`))) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(title = "Histogram logaritma sati slušanja dnevno",
       x = "log(Broj sati dnevno)",
       y = "Broj ispitanika") +
  theme_minimal()
```



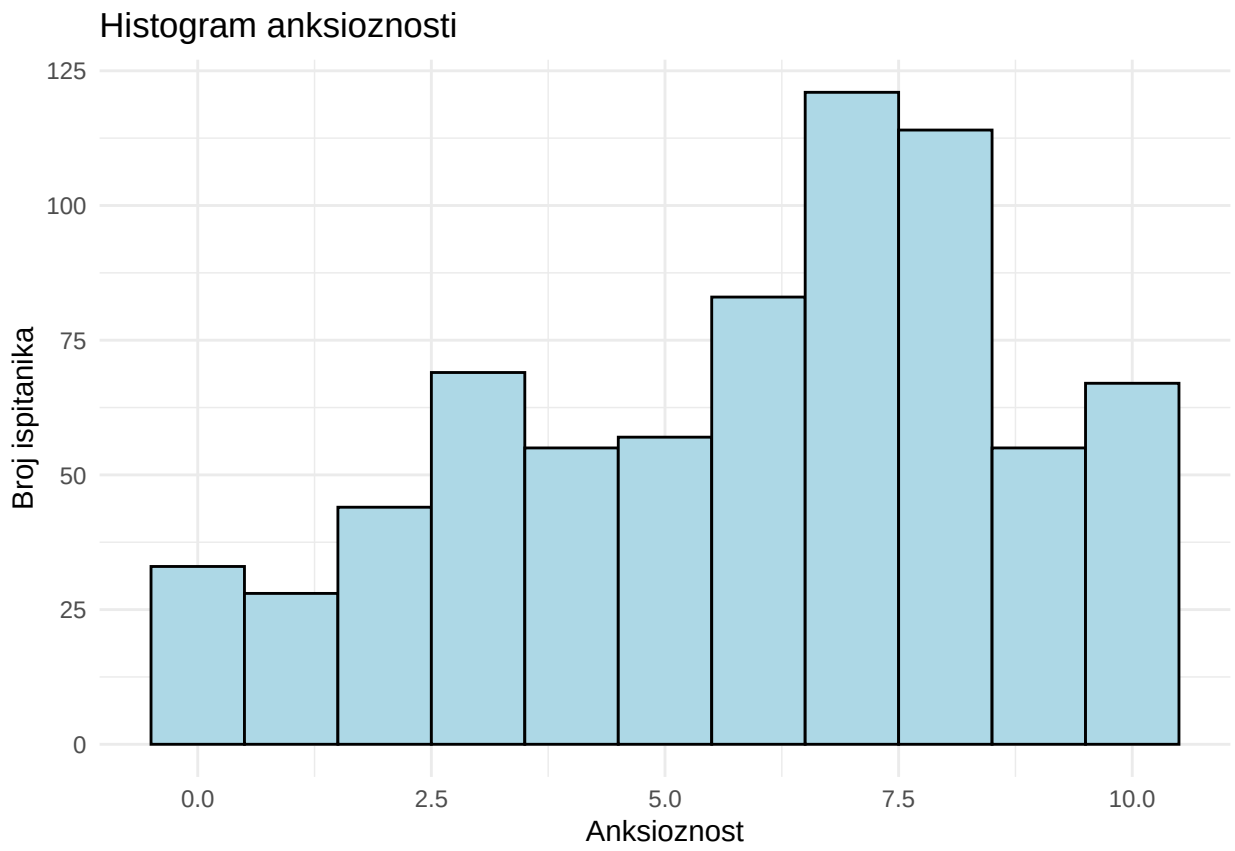
Iz tablice ispod svih histograma, vidimo da i dalje nisu normalno distribuirani. Osim i pomoću histograma i testa, možemo to vidjeti i qqplot-om koji ćemo provesti samo za varijablu `Hours per day` radi preglednosti.

```
qqnorm(music_data$`Hours per day`)
qqline(music_data$`Hours per day`, col = "steelblue", lwd = 2)
```

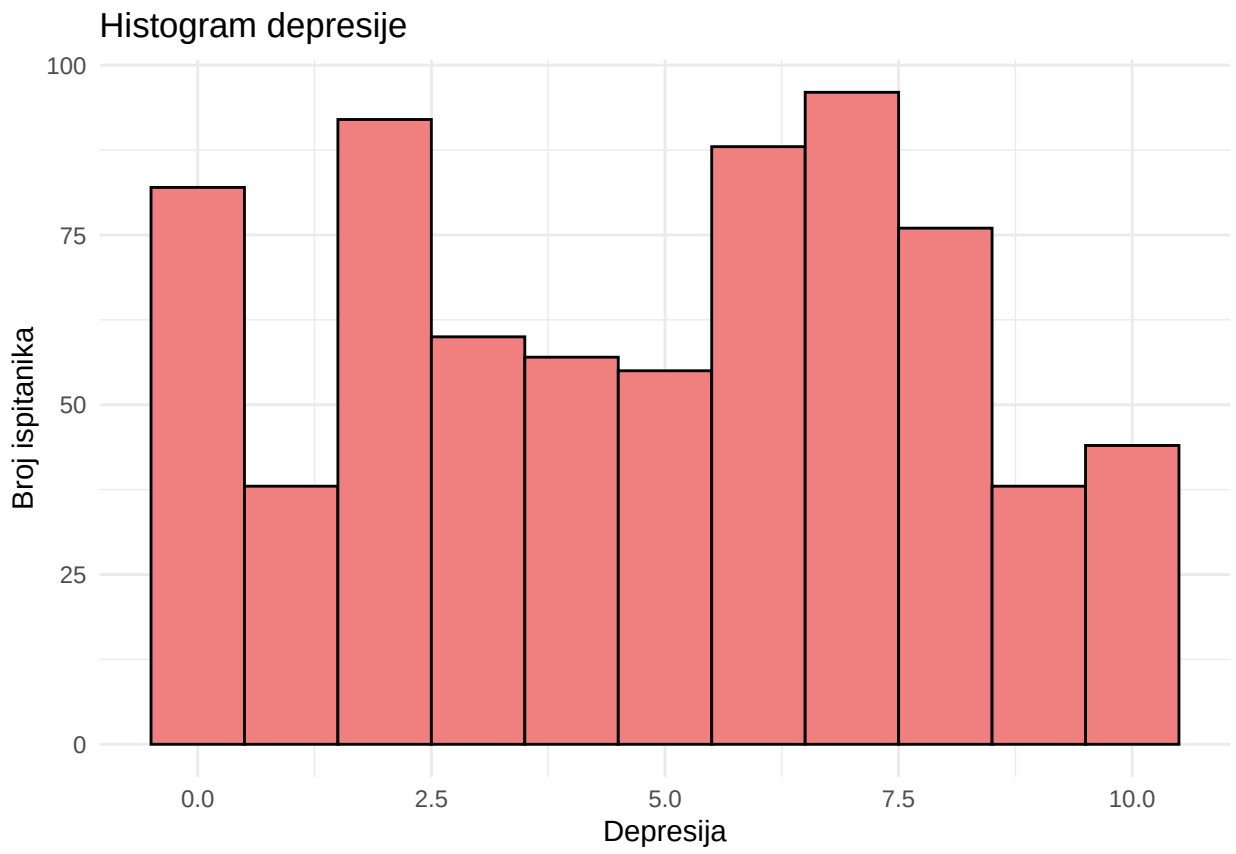


Ovdje opet vidimo da podaci nisu normalno distribuirani. Sada ćemo provjeriti normalnost distribucije za preostale varijable.

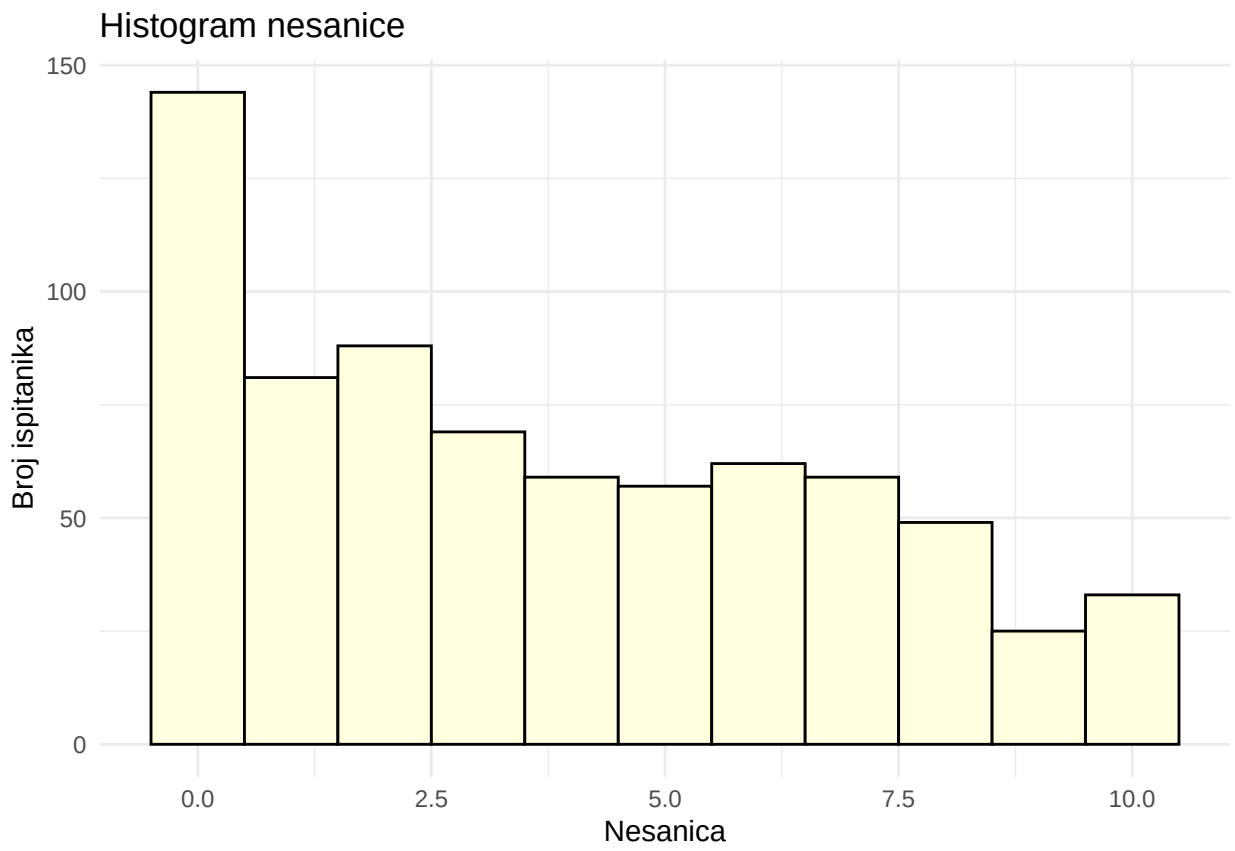
```
music_data %>%
  ggplot(aes(x = Anxiety)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(title = "Histogram anksioznosti",
       x = "Anksioznost",
       y = "Broj ispitanika") +
  theme_minimal()
```



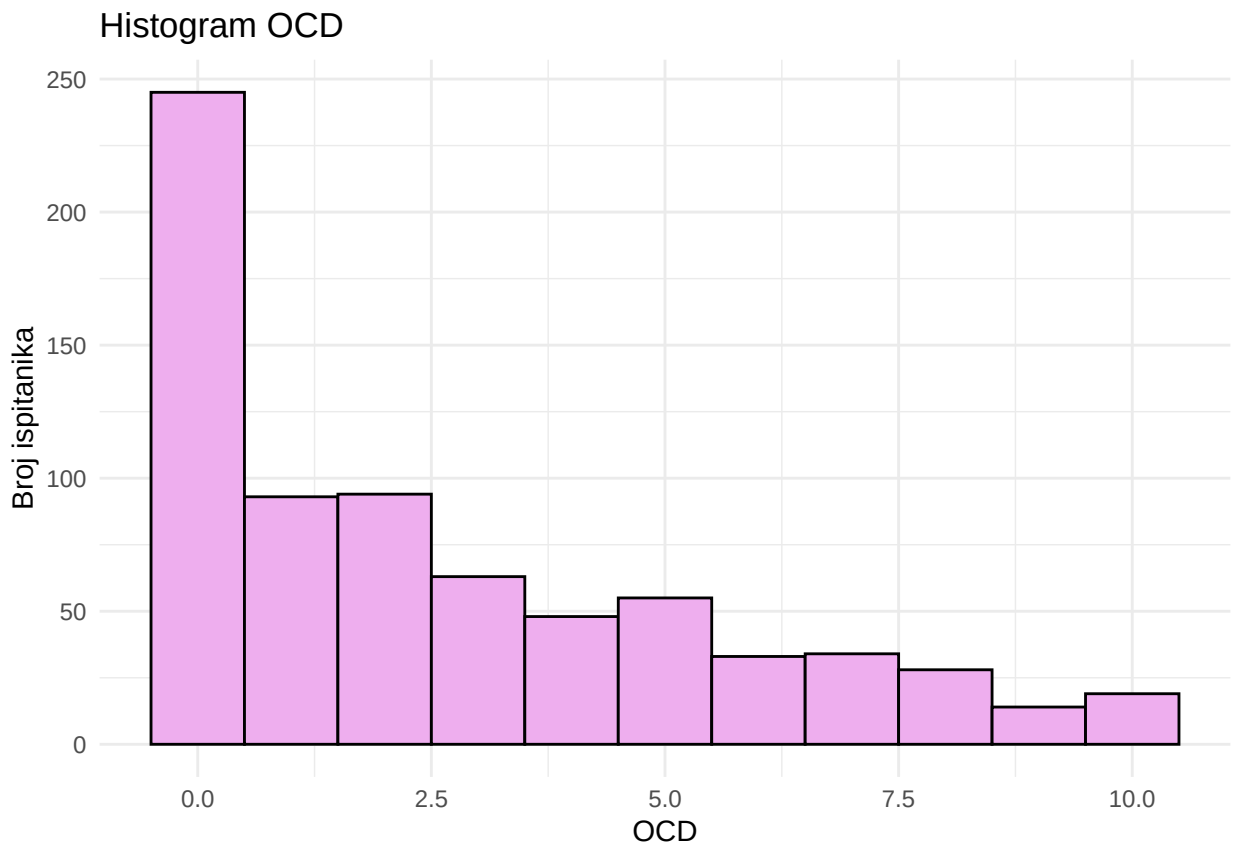

```
music_data %>%
  ggplot(aes(x = Depression)) +
  geom_histogram(binwidth = 1, fill = "lightcoral", color = "black") +
  labs(title = "Histogram depresije",
       x = "Depresija",
       y = "Broj ispitanika") +
  theme_minimal()
```




```
music_data %>%
  ggplot(aes(x = Insomnia)) +
  geom_histogram(binwidth = 1, fill = "lightyellow", color = "black") +
  labs(title = "Histogram nesanice",
       x = "Nesanica",
       y = "Broj ispitanika") +
  theme_minimal()
```




```
music_data %>%
  ggplot(aes(x = OCD)) +
  geom_histogram(binwidth = 1, fill = "plum2", color = "black") +
  labs(title = "Histogram OCD",
       x = "OCD",
       y = "Broj ispitanika")+
  theme_minimal()
```




```
analyze_normality(music_data)
```

```
## # A tibble: 6 x 3
##   Varijabla          `p-vrijednost` Distribucija
##   <chr>                       <dbl> <chr>
## 1 Anxiety                  1.77e-45 Ne normalna
## 2 Depression               8.99e-30 Ne normalna
## 3 Hours per day            6.29e-83 Ne normalna
## 4 Insomnia                 3.52e-42 Ne normalna
## 5 OCD                      3.78e-68 Ne normalna
## 6 log_Hours per day        3.56e-33 Ne normalna
```

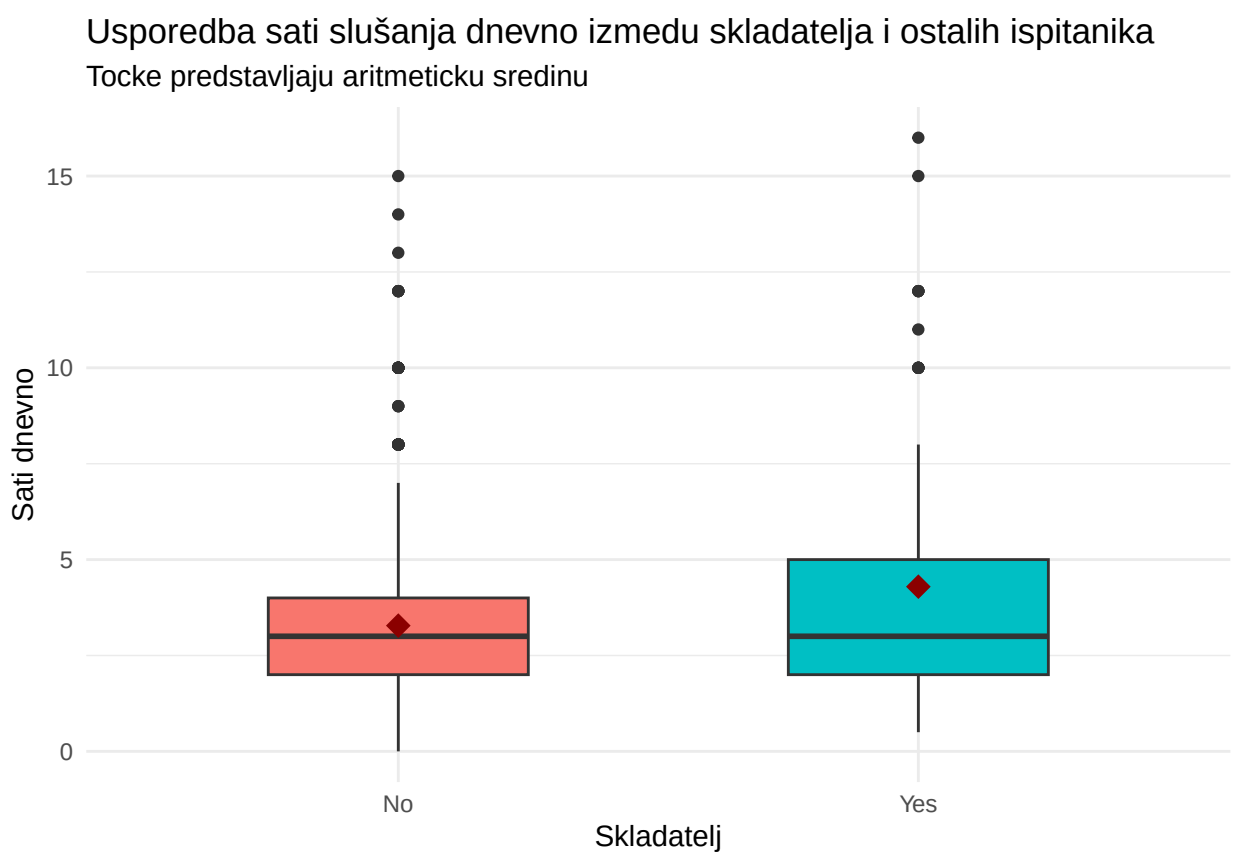
Zaključak: Na razini značajnosti $\alpha = 0.05$, odbacujemo nultu hipotezu za sve numeričke varijable jer p-vrijednosti testa su manje od 0.05. To znači da podaci **ne dolaze iz normalne distribucije**.

1. Slušaju li skladatelji više glazbe dnevno od drugih?

Izvdojimo podatke slušanja glazbe dnevno po tome jesu li ispitanici skladatelji ili ne i vizualizirajmo ih boxplotom.

```
music_data %>%
  ggplot(aes(x = Composer, y = `Hours per day`, fill = Composer)) +
  geom_boxplot(width=0.5) +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 4, color="darkred") +
```

```
  labs(title = "Usporedba sati slušanja dnevno između skladatelja i ostalih ispitanika",
       x = "Skladatelj",
       y = "Sati dnevno",
       subtitle = "Točke predstavljaju aritmetičku sredinu" ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Iz boxplota možemo zaključiti da skladatelji u prosjeku slušaju više glazbe dnevno od ostalih ispitanika. S obzirom na to da podaci nisu normalno distribuirani, provjerimo ovu tvrdnju neparametarskim Wilcoxonovim rank-sum testom. **Wilcoxonov rank-sum test** koristi se za uspoređivanje dva nezavisna uzorka. Nulta hipoteza testa je da nema razlike u satima slušanja dnevno između skladatelja i ostalih ispitanika, a test provodimo na razini značajnosti $\alpha = 0.05$.

```
composers <- music_data %>%
  filter(Composer == "Yes")

non_composers <- music_data %>%
  filter(Composer == "No")

wilcox.test(composers$`Hours per day`, non_composers$`Hours per day`, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  composers$`Hours per day` and non_composers$`Hours per day`
```

```
## W = 45120, p-value = 0.0001034
## alternative hypothesis: true location shift is greater than 0
```

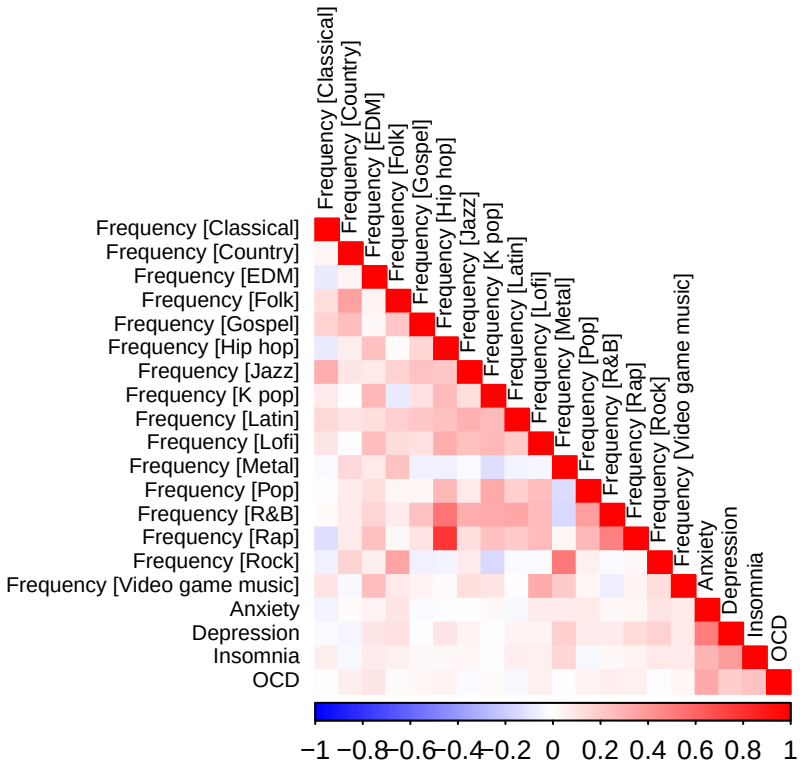
Zaključak: Prema rezultatima Wilcoxonovog rank-sum testa, odbacujemo nultu hipotezu na razini značajnosti $\alpha = 0.05$ u korist alternativne hipoteze. To znači da **skladatelji u prosjeku slušaju više glazbe dnevno od ostalih ispitanika**.

2. Korelacije izmedu frekvencija slušanja različitih glazbenih žanrova i samoprocijenjenih razina mentalnih poremećaja

Cilj ove analize je ispitati povezanost između učestalosti slušanja različitih glazbenih žanrova i samoprocijenjenih razina mentalnih poremećaja (anksioznost, depresija, nesanica, OCD). Korelacije nam pomažu identificirati obrasce u podacima, primjerice, je li povećana učestalost slušanja određenog žanra povezana s višim ili nižim razinama mentalnih poremećaja. Budući da podaci nisu normalno distribuirani, koristimo **Spearmanovu rang korelaciju**. Korelacija se izražava koeficijentom koji može imati vrijednosti od -1 (negativna korelacija) do +1 (pozitivna korelacija), dok p-vrijednost pokazuje je li korelacija statistički značajna.

Na kraju ćemo posebno analizirati povezanost između učestalosti slušanja metal žanra i razine depresije kako bismo provjerili postoji li značajna korelacija između tih varijabli.

```
# Odabiremo samo potrebne varijable
columns_of_interest <- c("Frequency [Classical]", "Frequency [Country]", "Frequency [EDM]", "Frequency
data_subset <- music_data[, columns_of_interest]

# Mapiranje, pretvorba kategorijskih frekvencija u numeričke
convert_frequency <- function(frequency) {
  case_when(
    frequency == "Never" ~ 0,
    frequency == "Rarely" ~ 1,
    frequency == "Sometimes" ~ 2,
    frequency == "Very frequently" ~ 3,
    TRUE ~ NA_real_
  )
}

data_subset <- data_subset %>%
  mutate(across(starts_with("Frequency"), convert_frequency))


# Izračun korelacija - Spearman
correlation_matrix <- cor(data_subset, method = "spearman", use = "pairwise.complete.obs")

# Vizualizacija
corrplot(correlation_matrix, method = "color", type = "lower", tl.col = "black",
         tl.cex = 0.7, col = colorRampPalette(c("blue", "white", "red"))(200))
```




```
metal_depression_test <- cor.test(data_subset$`Frequency [Metal]`, data_subset$Depression, method = "sp
```

```
## Warning in cor.test.default(data_subset$`Frequency [Metal]`,
## data_subset$Depression, : Cannot compute exact p-value with ties
```

```
print(metal_depression_test)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data_subset$`Frequency [Metal]` and data_subset$Depression
## S = 52255950, p-value = 9.631e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.180634
```

Zaključak: Rezultati Spearmanove korelacije pokazuju slab pozitivni odnos (rho = 0.18) između učestalosti slušanja metal glazbe i samoprocijenjene razine depresije. To znači da je moguće da ispitanici koji češće slušaju metal glazbu imaju nešto višu razinu depresije. P-vrijednost iznosi 9.631e-07, što ukazuje na to da je korelacija statistički značajna. Ovaj rezultat sugerira da je **odnos između slušanja metal glazbe i depresije vjerojatno stvaran** i nije posljedica slučajnosti.

3. Postoji li povezanost između korištenog servisa i sklonosti istraživanju

Kako bismo provjerili ovu tvrdnju provodimo **test nezavisnosti** χ^2. Nulta hipoteza testa je da nema povezanosti između primarnog streaming servisa i sklonosti istraživanju, a test provodimo na razini značajnosti $\alpha = 0.05$. Umjesto da izbacujemo streaming servis `Pandora`, grupirali smo ga zajedno s `Other` stavkom kako bismo zadovoljili pretpostavku testa da su frekvencije podataka ≥ 5. Također, izbacili smo `None` jer su to ispitanici koji ne koriste streaming servise.

```
music_data %>%
  mutate(`Primary streaming service` = ifelse(`Primary streaming service` == "Pandora", "Other", `Prima
  filter(`Primary streaming service` != "None") %>%
  with(table(`Primary streaming service`, Exploratory)) -> tbl

tbl
```

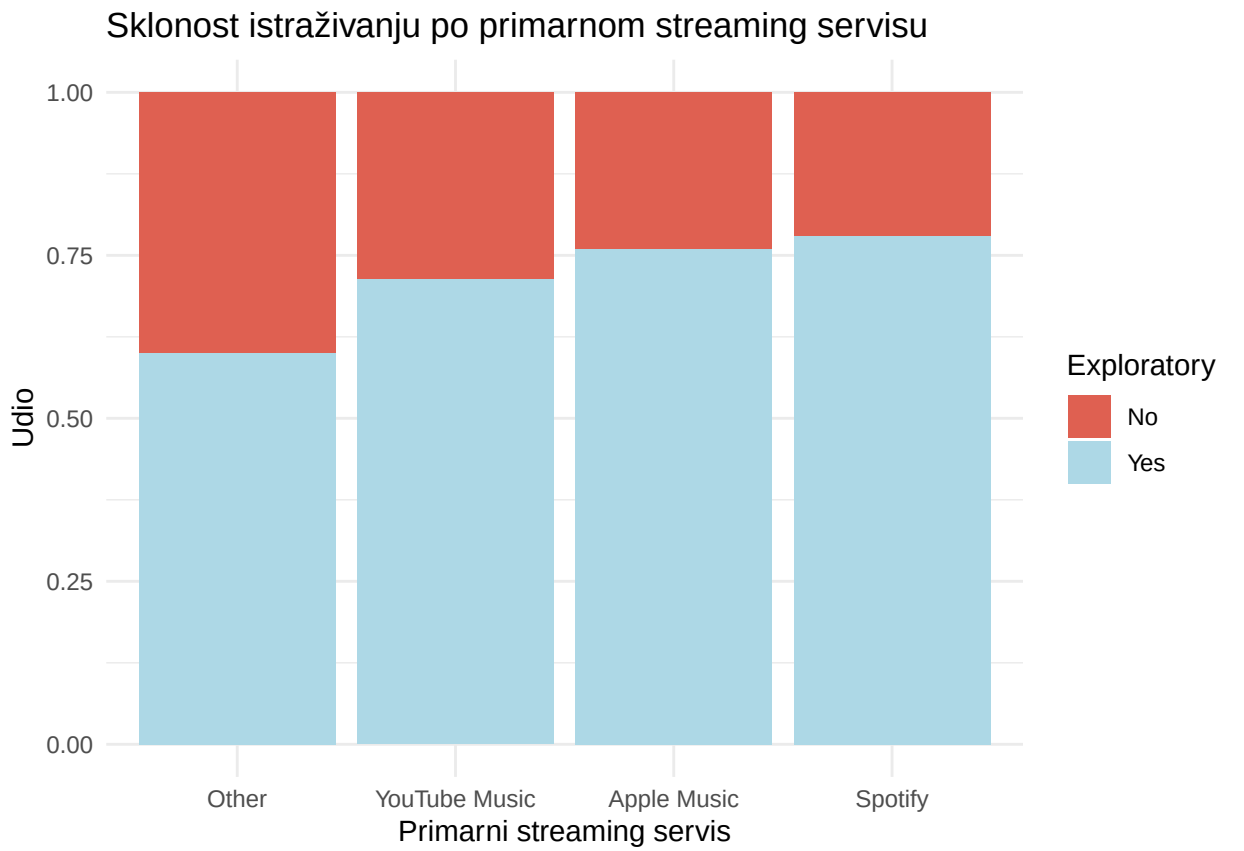
```
##                          Exploratory
## Primary streaming service  No Yes
##             Apple Music    12  38
##             Other          24  36
##             Spotify       100 353
##             YouTube Music  27  67
```

```
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 10.027, df = 3, p-value = 0.01834
```

Na razini značajnosti 0.05, odbacujemo nultu hipotezu u korist alternativne hipoteze. To znači da postoji povezanost između primarnog streaming servisa i sklonosti istraživanju. Naime da smo koristili razinu značajnosti 0.01, ne bismo odbacili nultu hipotezu.

```
music_data %>%
    mutate(`Primary streaming service` = ifelse(`Primary streaming service` == "Pandora", "Other", `Pri
  filter(`Primary streaming service` != "None") %>%
  ggplot(aes(x = reorder(`Primary streaming service`, Exploratory == "Yes", FUN = mean),
             fill = Exploratory)) +
  geom_bar(position = "fill") +
  labs(title = "Sklonost istraživanju po primarnom streaming servisu",
       x = "Primarni streaming servis",
       y = "Udio") +
  scale_fill_manual(values = c("No" = "#DF6051", "Yes" = "lightblue")) +
  theme_minimal()
```



Zaključak: Prema rezultatima χ^2 testa, odbacili smo nultu hipotezu u korist alternativne. To znači da **postoji povezanost između primarnog streaming servisa i sklonosti istraživanju**. Vizualizacijom iznad možemo primijetiti da korisnici Spotifyja i Apple Musica imaju veću sklonost istraživanju u odnosu na korisnike drugih servisa.

4. Razlikuje li se prosječni broj sati slušanja glazbe značajno među korisnicima ovisno o njihovim omiljenim žanrovima?

ANOVA (engl. *ANalysis Of VAriance*) je metoda kojom testiramo sredine više populacija. U ovom slučaju, testiramo razlike u prosječnom broju sati slušanja glazbe dnevno među korisnicima ovisno o njihovim omiljenim žanrovima. Nulta hipoteza testa je da nema razlike u prosječnom broju sati slušanja dnevno među korisnicima omiljenih žanrova, a test provodimo na razini značajnosti $\alpha = 0.05$. Radi distribucije podataka, koristimo **Kruskal-Wallisov test** koji je neparametarska verzija ANOVA-e.

Za početak, prebrojavamo koliko ispitanika preferira svaki žanr.

```
music_data %>%
  group_by(`Fav genre`) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 16 x 2
##    `Fav genre`      count
##    <chr>            <int>
##  1 Rock               185
```

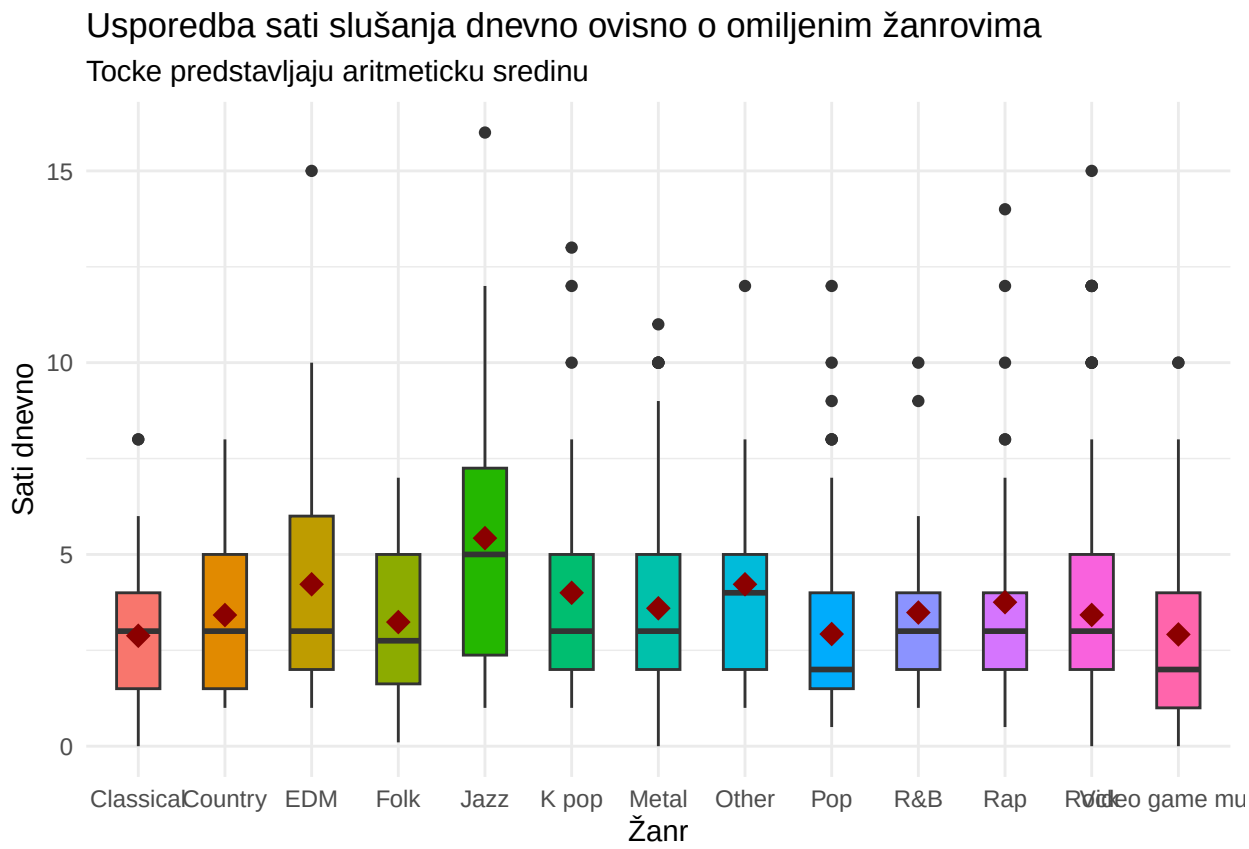
```
##  2 Pop                113
##  3 Metal               88
##  4 Classical           52
##  5 Video game music    44
##  6 EDM                 36
##  7 Hip hop             35
##  8 R&B                 34
##  9 Folk                30
## 10 K pop               26
## 11 Country             25
## 12 Jazz                20
## 13 Rap                 20
## 14 Lofi                10
## 15 Gospel               5
## 16 Latin                3
```

Radi preglednosti, grupirat ćemo žanrove s manje od 10 ispitanika u kategoriju `Other`. Također, spojit ćemo hip-hop i rap jer su često povezani žanrovi.

```
music_data <- music_data %>%
  mutate(`Fav genre` = ifelse(`Fav genre` == "Latin" | `Fav genre` == "Lofi" | `Fav genre` == "Gospel",
  mutate(`Fav genre` = ifelse(`Fav genre` == "Hip hop", "Rap", `Fav genre`))
```

Zatim vizualiziramo boxplotom razlike u prosječnom broju sati slušanja dnevno među korisnicima omiljenih žanrova.

```
music_data %>%
  ggplot(aes(x = `Fav genre`, y = `Hours per day`, fill = `Fav genre`)) +
  geom_boxplot(width=0.5) +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 4, color="darkred") +
  labs(title = "Usporedba sati slušanja dnevno ovisno o omiljenim žanrovima",
       x = "Žanr",
       y = "Sati dnevno",
       subtitle = "Točke predstavljaju aritmetičku sredinu" ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Provodimo **test homogenosti varijanci** kako bismo provjerili pretpostavku o jednakim varijancama među skupinama. Nulta hipoteza testa je da su varijance jednake, a test provodimo na razini značajnosti $\alpha = 0.05$.

```
music_data$`Fav genre` <- as.factor(music_data$`Fav genre`)

leveneTest(`Hours per day` ~ `Fav genre`, data = music_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  12  1.5491 0.1019
##       713
```

```
music_data$`Fav genre` <- as.factor(music_data$`Fav genre`)
bartlett.test(`Hours per day` ~ `Fav genre`, data = music_data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Hours per day by Fav genre
## Bartlett's K-squared = 39.544, df = 12, p-value = 8.558e-05
```

Vidimo da je **Levene test** bolji odabir jer naši podaci ne podliježu normalnoj razdiobi, a **Bartlettov test** je osjetljiv na ne-normalnu razdiobu.

Na razini značajnosti 0.05, ne odbacujemo nultu hipotezu. To znači da su varijance jednake među skupinama. S obzirom na to, provodimo Kruskal-Wallisov test.

```r
kruskal.test(`Hours per day` ~ `Fav genre`, data = music_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Hours per day by Fav genre
## Kruskal-Wallis chi-squared = 23.413, df = 12, p-value = 0.02441
```

Zaključak: Na razini značajnosti 0.05, odbacujemo nultu hipotezu u korist alternativne. To znači da **postoji značajna razlika** u prosječnom broju sati slušanja dnevno ovisno o omiljenim žanrovima korisnika.

5. Može li se iz zadanih podataka predvidjeti dob ispitanika?

U ovom zadatku koristiti ćemo linearnu regresiju kako bismo pokušali predvidjeti dob ispitanika. Glavne pretpostavke modela linearne regresije su:

Linearnost

- Veza između nezavisnih i zavisne varijable mora biti linearna
- Može se prikazati jednadžbom: $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon$

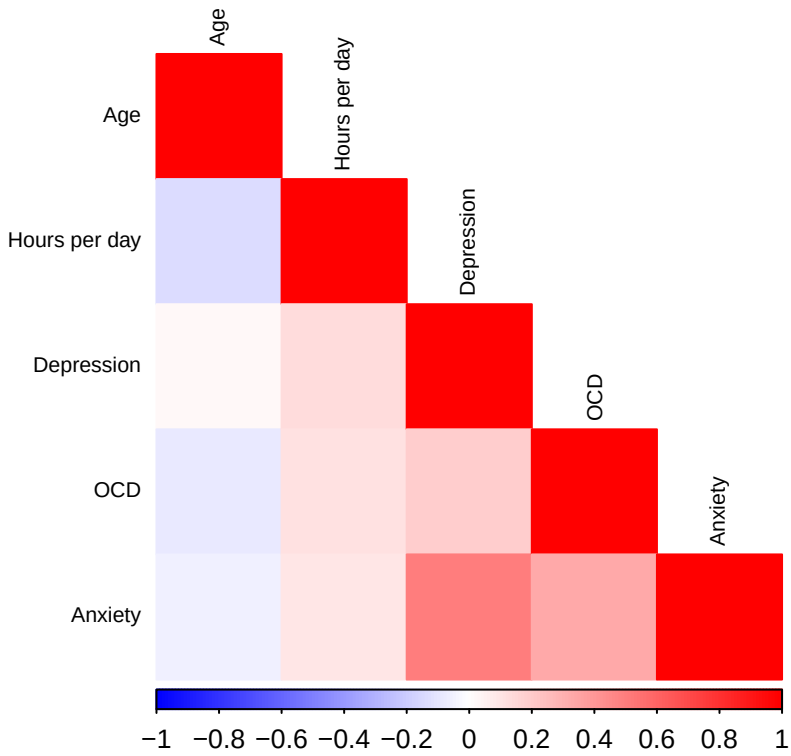
Normalnost reziduala

- Reziduali trebaju pratiti normalnu distribuciju
- Provjerava se Q-Q plotom / histogramom i statističkim testovima

Posljedice kršenja pretpostavki: Pristrasne procjene parametara, nepouzdani rezultati modela, rezultati nisu značajni

Napraviti ćemo korelacijske tablice da vidimo koje varijable bi mogle biti značajne za predviđanje dobi ispitanika. Nakon toga, provest ćemo linearnu regresiju.

```r
numericvars <- music_data %>%
  select(Age, `Hours per day`, `Depression`, `OCD`, `Anxiety`)
correlation_matrix <- cor(numericvars, method = "spearman", use = "pairwise.complete.obs")


corrplot(correlation_matrix, method = "color", type = "lower", tl.col = "black",
         tl.cex = 0.7, col = colorRampPalette(c("blue", "white", "red"))(200))
```

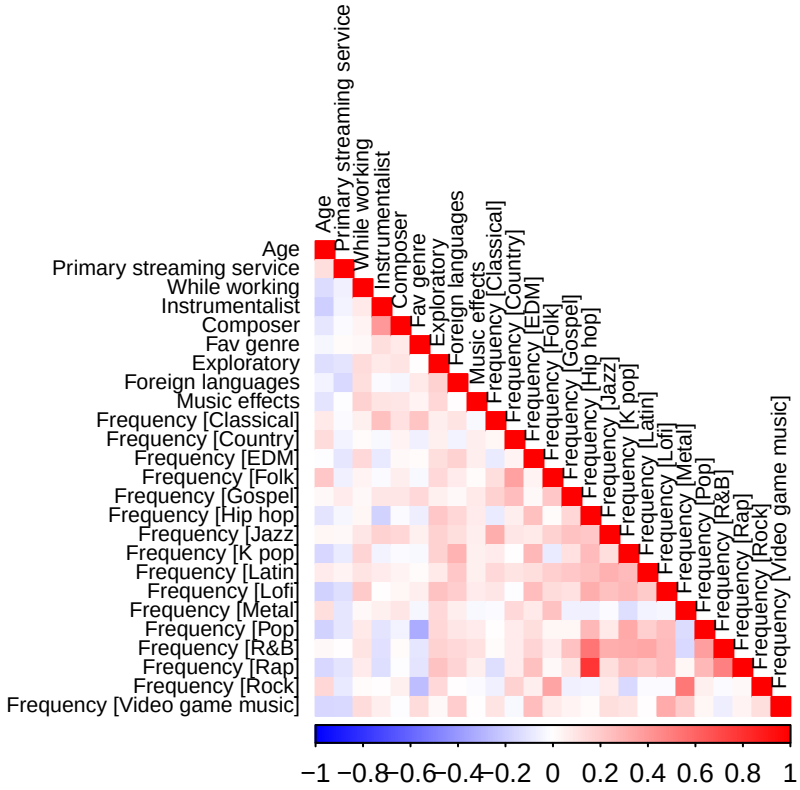


Definiramo funkcije za pretvorbu kategoričkih varijabli u dummy varijable kako bismo ih mogli koristiti u regresiji i korelacijskom testiranju.

```
convert_frequency <- function(x) {
  freq_levels <- c("Never" = 0, "Rarely" = 1, "Sometimes" = 2, "Very frequently" = 3)
  as.numeric(factor(x, levels = names(freq_levels))) - 1
}

convert_binary <- function(x) {
  as.numeric(factor(x, levels = c("No", "Yes")))
}

convert_music_effects <- function(x) {
  as.numeric(factor(x, levels = c("Worsen","No effect", "Improve"))) - 2
}

convert_primary_streaming_service <- function(x) {
  as.numeric(factor(x, levels = c("Spotify", "Apple Music", "Youtube Music", "Pandora", "Other")))
}

convert_genre <- function(x) {
  as.numeric(factor(x, levels = c("Pop", "Rock", "Rap", "Hip hop", "Country", "Jazz", "Metal", "EDM", "
}
```

Radimo korelacijsku tablicu između `Age` i kategoričkih varijabli.

```r
corr_data <- music_data %>%
  select(`Age`, `Primary streaming service`, `While working`, Instrumentalist,
         Composer, `Fav genre`, Exploratory, `Foreign languages`, `Music effects`, starts_with("Frequency
  mutate(
    across(starts_with("Frequency"), convert_frequency),
    across(c(`While working`, Instrumentalist, Composer, Exploratory, `Foreign languages`), convert_bina
   `Primary streaming service` = convert_primary_streaming_service(`Primary streaming service`),
   `Fav genre` = convert_genre(`Fav genre`),
   `Music effects` = convert_music_effects(`Music effects`)
 )

correlation_matrix <- cor(corr_data, method = "spearman", use = "pairwise.complete.obs")
corrplot(correlation_matrix, method = "color", type = "lower",
         tl.col = "black", tl.cex = 0.7,
         col = colorRampPalette(c("blue", "white", "red"))(200))
```




```r
age_correlations <- abs(correlation_matrix["Age", ])
age_correlations <- age_correlations[order(age_correlations, decreasing = TRUE)]

age_correlations_df <- data.frame(
  Variable = names(age_correlations),
  Correlation = as.numeric(age_correlations)
)
print(age_correlations_df)
```

```
##                         Variable Correlation
## 1                            Age 1.000000000
## 2              Frequency [Folk] 0.212890175
## 3               Instrumentalist 0.185558916
## 4              Frequency [Lofi] 0.174272556
## 5               Frequency [Pop] 0.170934299
## 6              Frequency [Rock] 0.156424543
## 7  Frequency [Video game music] 0.156116494
## 8              Frequency [K pop] 0.153584777
## 9                Frequency [Rap] 0.151701642
## 10                While working 0.137030300
## 11           Frequency [Country] 0.135047113
## 12    Primary streaming service 0.129027653
## 13                  Exploratory 0.123611059
## 14             Frequency [Metal] 0.120217893
## 15                Music effects 0.107929021
## 16           Frequency [Hip hop] 0.106710614
## 17                     Composer 0.093744764
## 18         Frequency [Classical] 0.083365485
## 19             Frequency [Latin] 0.078530412
## 20             Foreign languages 0.041318914
## 21                     Fav genre 0.039444195
## 22              Frequency [Jazz] 0.032764175
## 23            Frequency [Gospel] 0.023864613
## 24               Frequency [R&B] 0.021689884
## 25               Frequency [EDM] 0.005266463
```

Prema korelacijskim tablicama uzeti ćemo varijable koje imaju najveću **apsolutnu vrijednost** korelacije s varijablom `Age` za linearnu regresiju. Među numeričkim varijablama to su `Hours per day` i `OCD` dok su kod kategoričkih varijabli to frekvencije slušanja glazbe. Od kategoričkih izdvojiti ćemo slijedeće: `Frequency [Folk]`, `Frequency[Pop]`,`Frequency [Country]`, `Instrumentalist`, `While Working` i `Primary streaming service`.

Provedimo prvo par jednostavnih linearnih regresija za varijable `OCD` i `Hours per day` kako bismo vidjeli jesu li te varijable značajne za model.

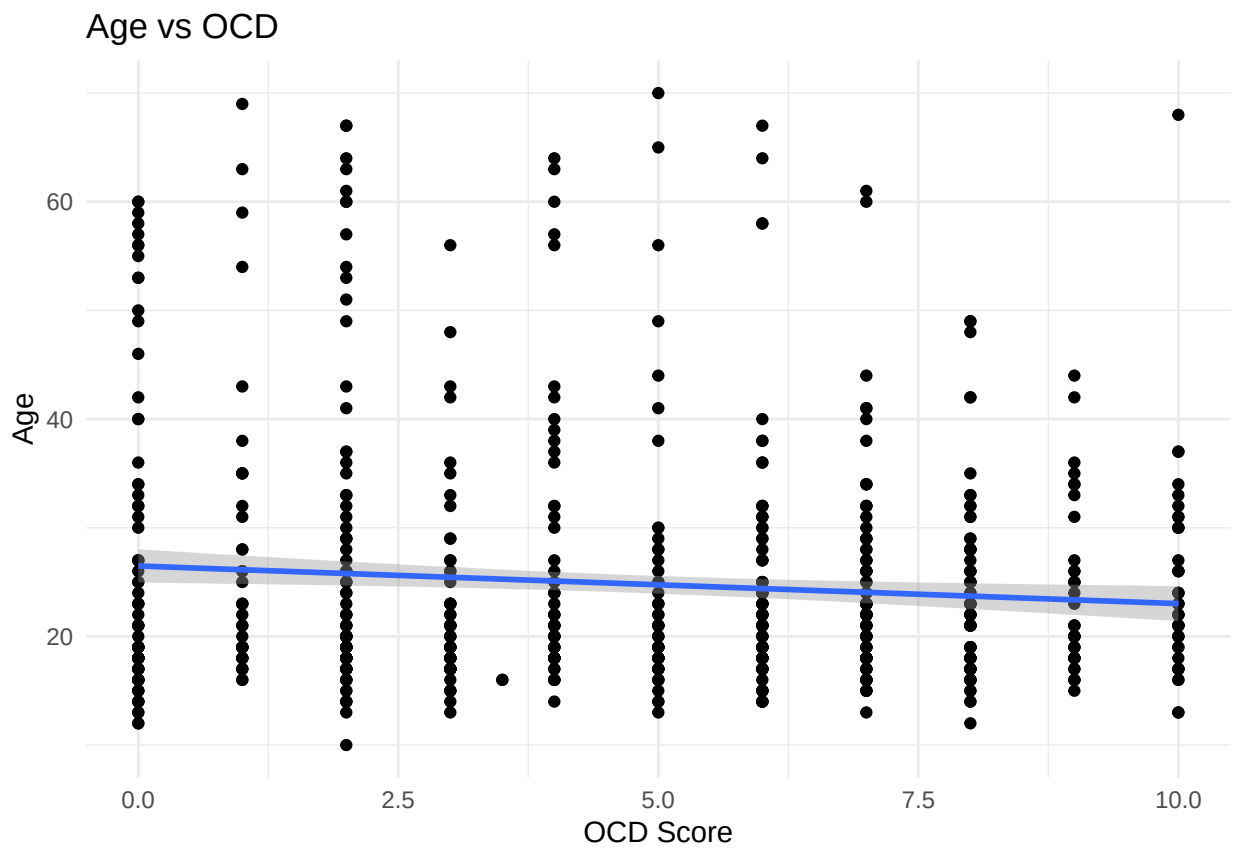
```
model <- lm(Age ~ OCD, data = music_data)

summary(model)
```

```
##
## Call:
## lm(formula = Age ~ OCD, data = music_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.566  -7.027  -3.644   2.798  44.434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.0270     0.5624  46.280  < 2e-16 ***
## OCD          -0.4611     0.1453  -3.173  0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

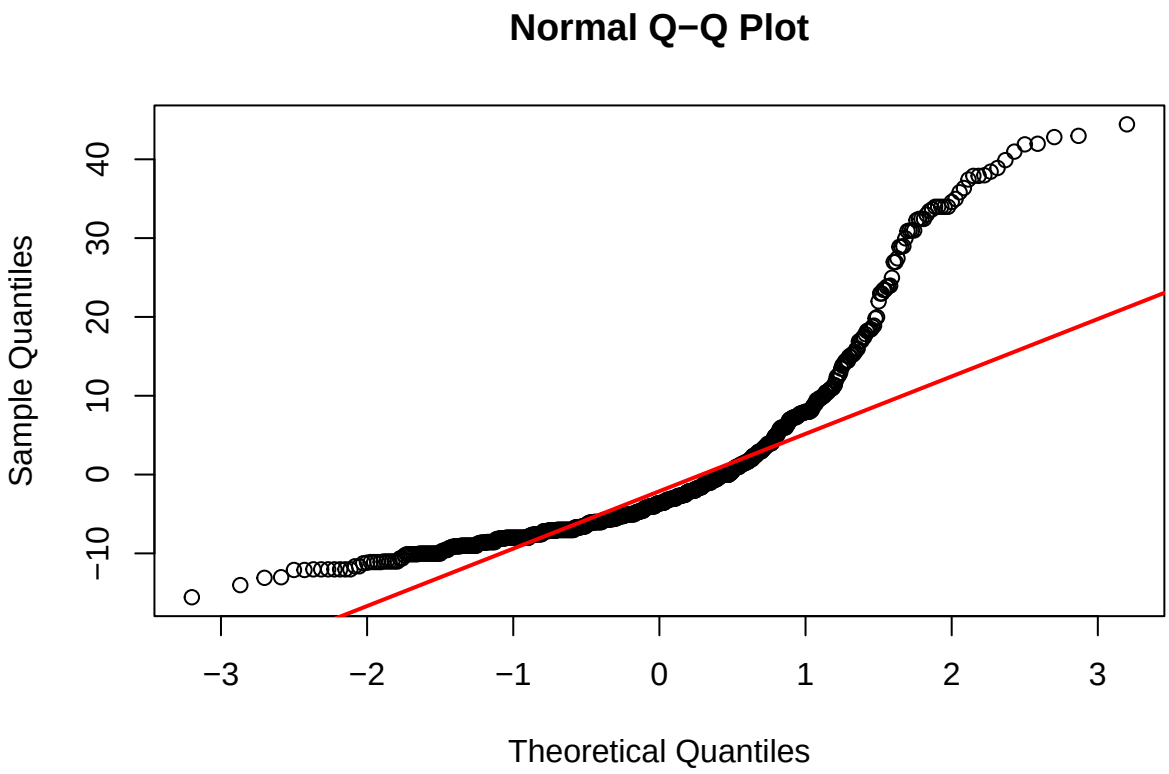
```
##
## Residual standard error: 11.09 on 724 degrees of freedom
## Multiple R-squared:  0.01372,    Adjusted R-squared:  0.01235
## F-statistic: 10.07 on 1 and 724 DF,  p-value: 0.001572
```

```r
ggplot(music_data, aes(x = Depression, y = Age)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_minimal() +
  labs(title = "Age vs OCD",
       x = "OCD Score",
       y = "Age")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
qqnorm(model$residuals)
qqline(model$residuals, col = "red", lwd = 2)
```




```
lillie.test(model$residuals)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  model$residuals
## D = 0.18417, p-value < 2.2e-16
```

Vidimo da je OCD statistički značajan no slabo objašnjava dob ispitanika (niski R^2). Također, reziduali ne podliježu normalnoj razdiobi što znači da to nije značajna varijabla za model. Sada ćemo provesti linearnu regresiju za varijablu `Hours per day`.

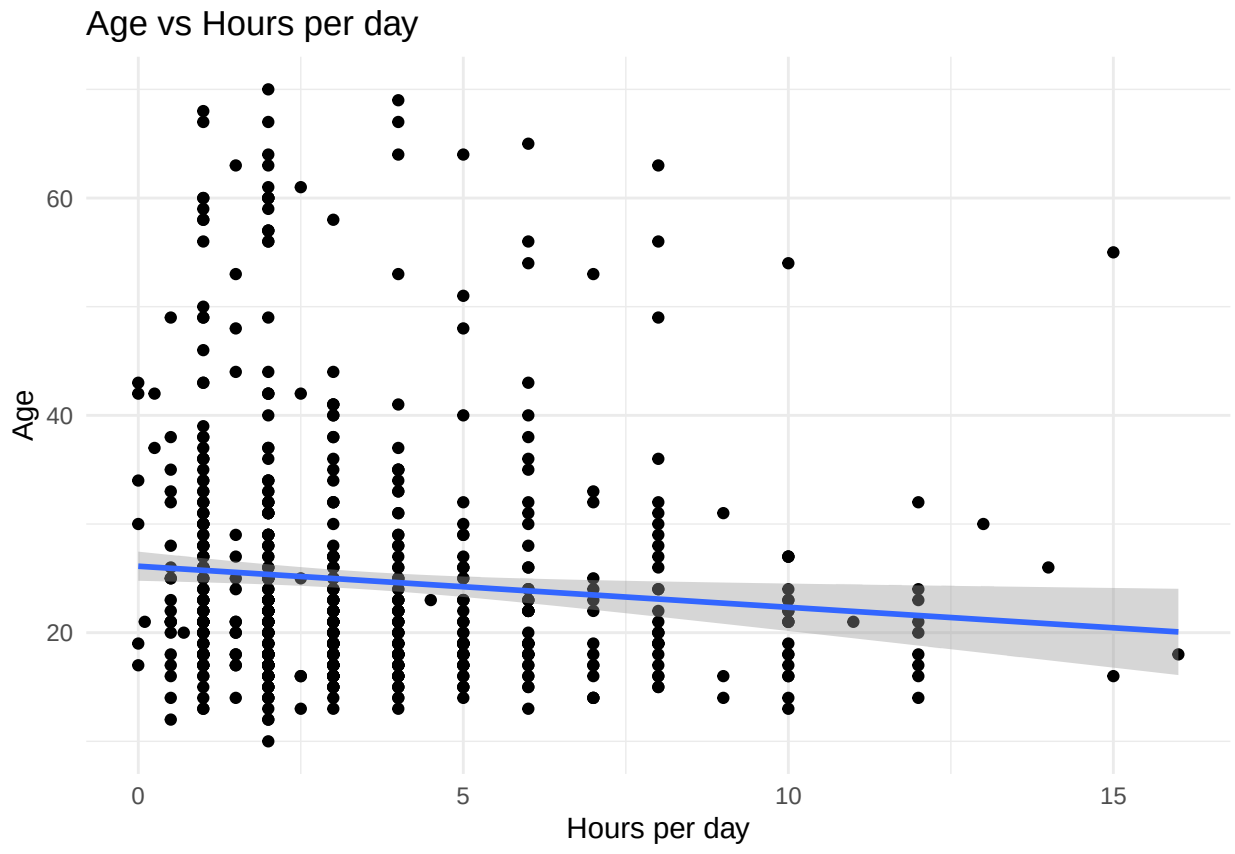
```
model <- lm(Age ~ `Hours per day`, data = music_data)

summary(model)
```

```
##
## Call:
## lm(formula = Age ~ `Hours per day`, data = music_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.360  -6.982  -3.982   2.585  44.640
##
```

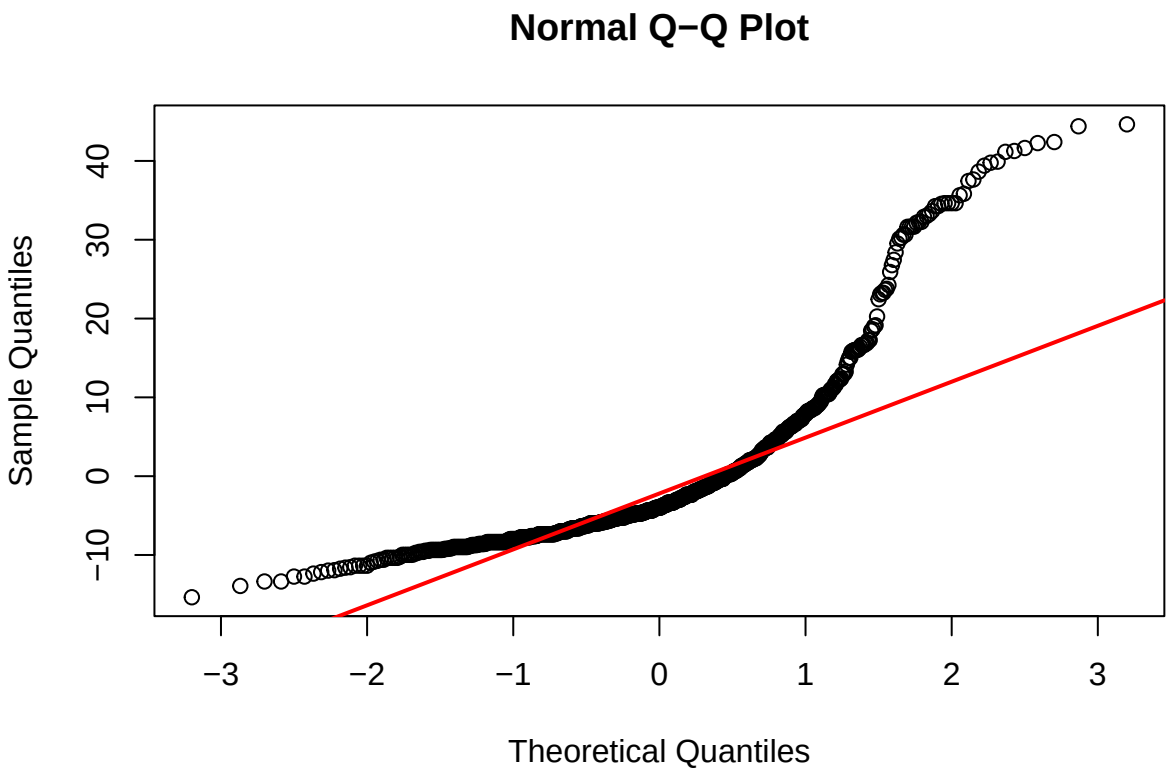
```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      26.1159     0.6834  38.216   <2e-16 ***
## `Hours per day`  -0.3779     0.1577  -2.396   0.0168 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.13 on 724 degrees of freedom
## Multiple R-squared:  0.007866,   Adjusted R-squared:  0.006496
## F-statistic:  5.74 on 1 and 724 DF,  p-value: 0.01683
```

```
ggplot(music_data, aes(x = `Hours per day`, y = Age)) +
  geom_point() +
  geom_smooth(method = "lm") + # 95% confidence interval
  theme_minimal() +
  labs(title = "Age vs Hours per day",
       x = "Hours per day",
       y = "Age")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
qqnorm(model$residuals)
qqline(model$residuals, col = "red", lwd = 2)
```




```
lillie.test(model$residuals)
```

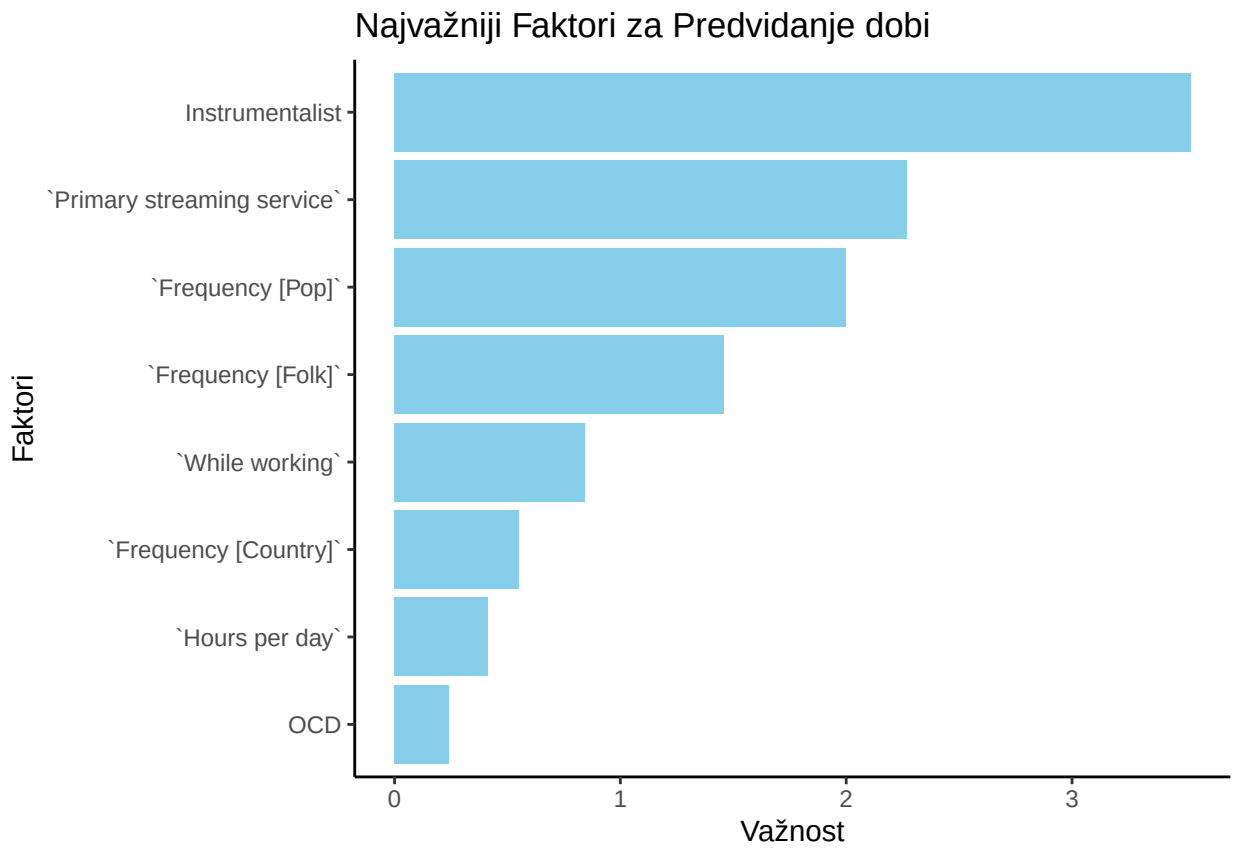
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  model$residuals
## D = 0.18151, p-value < 2.2e-16
```

Vidimo da je veza između `Hours per day` i `Age` statistički značajna no R^2 vrijednost je vrlo niska. Reziduali ne podliježu normalnoj razdiobi što dodatno znači da to nije značajna varijabla za model. Iako je statistički značajna veza, rezultati nam govore da postoje faktori koji bolje objašnjavaju varijancu podataka.

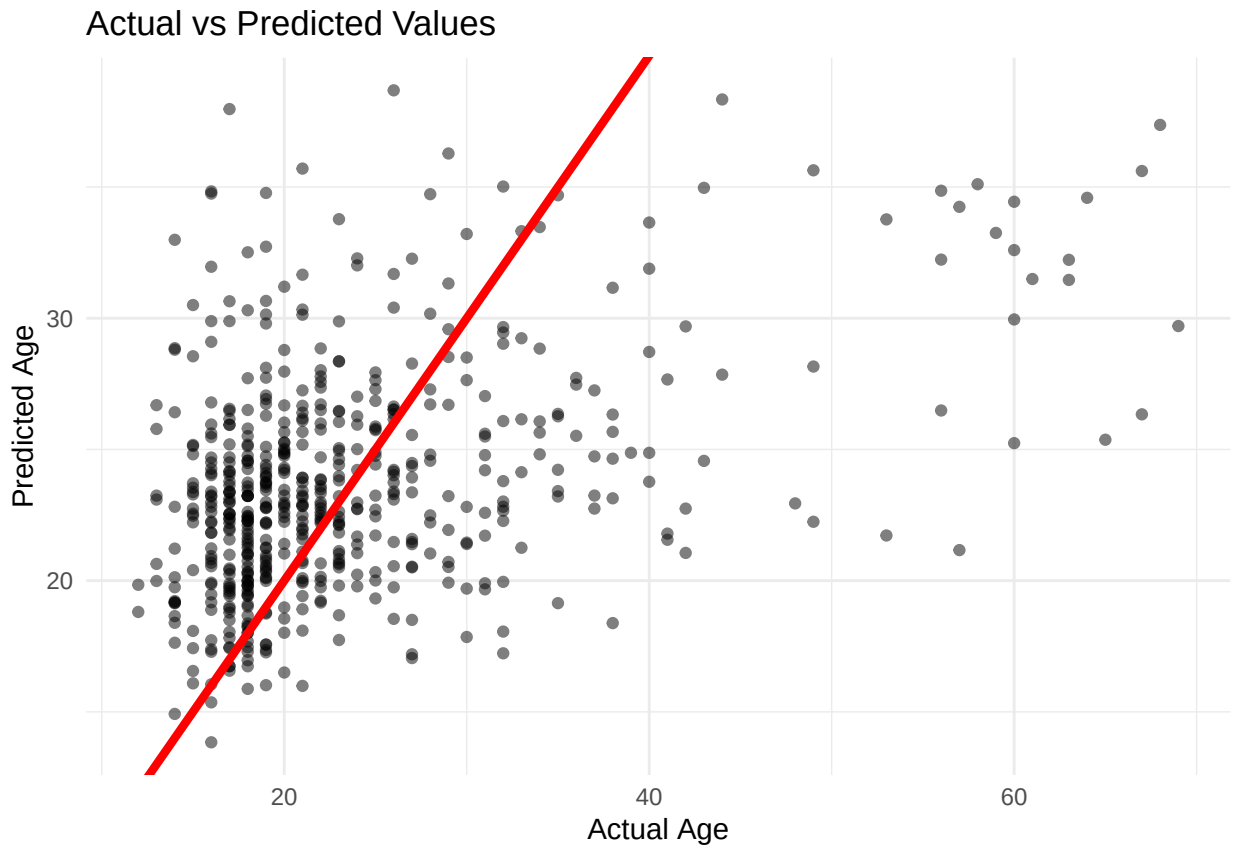
Probajmo sada napraviti višestruku regresiju koja uključuje sve varijable koje smo prije izvdojili.

```
chosen_data <- music_data %>%
  select(`Age`, `Hours per day`, `OCD`, `Primary streaming service`, `While working`, Instrumentalist,
  mutate(
    across(starts_with("Frequency"), convert_frequency),
    across(c(`While working`, Instrumentalist), convert_binary),
   `Primary streaming service` = convert_primary_streaming_service(`Primary streaming service`)
 )
```

```
results <- multiple_var_analysis(chosen_data)

print(results$top_10_plot)
```




```
print(results$prediction_plot)
```

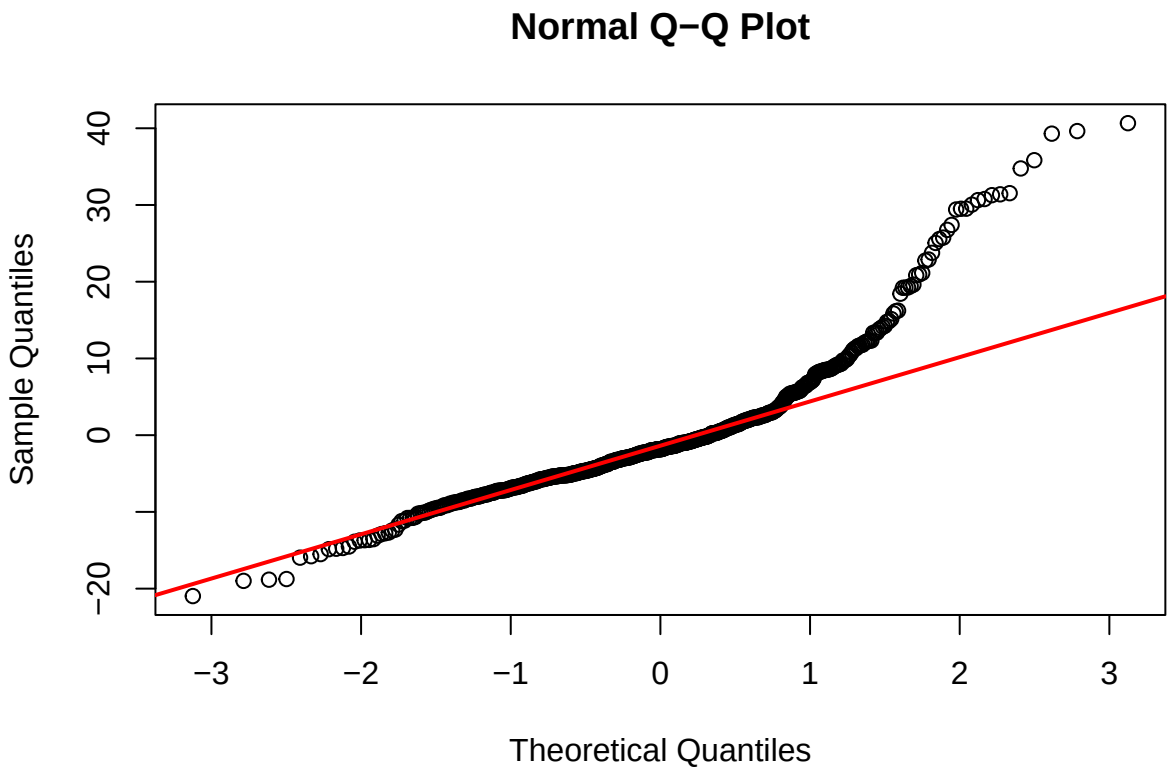



```
print(summary(results$model))
```

```
##
## Call:
## lm(formula = Age ~ ., data = chosen_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.970  -5.268  -1.829   2.516  40.662
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  30.9938     2.3519  13.178  < 2e-16 ***
## `Hours per day`              -0.4139     0.1618  -2.558 0.010806 *
## OCD                          -0.2411     0.1385  -1.741 0.082255 .
## `Primary streaming service`   2.2685     0.3288   6.900 1.43e-11 ***
## `While working`              -0.8416     1.0288  -0.818 0.413672
## Instrumentalist              -3.5250     0.8329  -4.232 2.71e-05 ***
## `Frequency [Folk]`            1.4602     0.4096   3.565 0.000395 ***
## `Frequency [Pop]`            -1.9990     0.4335  -4.611 4.97e-06 ***
## `Frequency [Country]`         0.5506     0.4519   1.218 0.223599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.167 on 552 degrees of freedom
```

```
## Multiple R-squared:  0.1877, Adjusted R-squared:  0.1759
## F-statistic: 15.94 on 8 and 552 DF,  p-value: < 2.2e-16
```

```
qqnorm(results$model$residuals)
qqline(results$model$residuals, col = "red", lwd = 2)
```



Prema rezultatima višestruke regresije koristeći varijable koje smo prije izdvojili, vidimo da model nije savršeno objašnjen tim odabirom varijabli. R^2 vrijednost je niska iako model bolje objašnjava dob ispitanika nego pojedinačne varijable. Prema qqplotu, reziduali većinom podliježu normalnoj razdiobi što znači da je model ima neku prediktivnu moć.

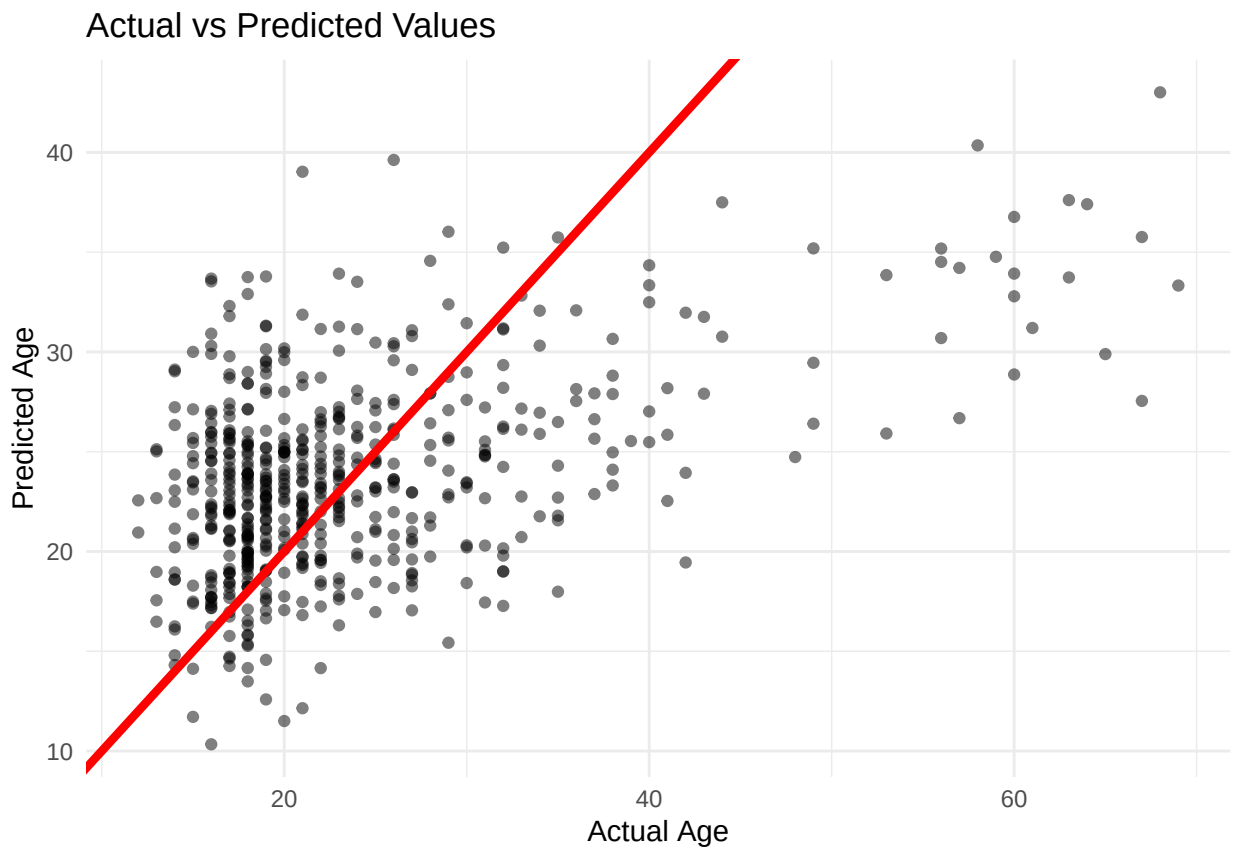
Probajmo poboljšati model odabirom drugih varijabli, ostavljajući varijable koje najbolje objašnjavaju dob ispitanika u prošloj regresiji: `Instrumentalist`, `Primary Streaming Service`, `Frequency [Pop]`, `Frequency [Folk]`, `Hours per day` i `OCD`.

Dodati ćemo slijedeće varijable koje bi mogle bolje objasniti dob ispitanika: `Exploratory`, `Frequency [Classical]`, `Depression`, `Frequency [R&B]` i `Frequency [Video game music]`.

```
chosen_data2 <- music_data %>%
  select(`Age`, `Primary streaming service`, Instrumentalist, `Frequency [Folk]`, `Frequency [Pop]`,`Hou
  mutate(
    across(starts_with("Frequency"), convert_frequency),
    across(c(Instrumentalist, `Exploratory`, `Composer`), convert_binary),
   `Primary streaming service` = convert_primary_streaming_service(`Primary streaming service`)
 )
```

```r
results2 <- multiple_var_analysis(chosen_data2)

print(results2$top_10_plot)
```


```r
print(results2$prediction_plot)
```

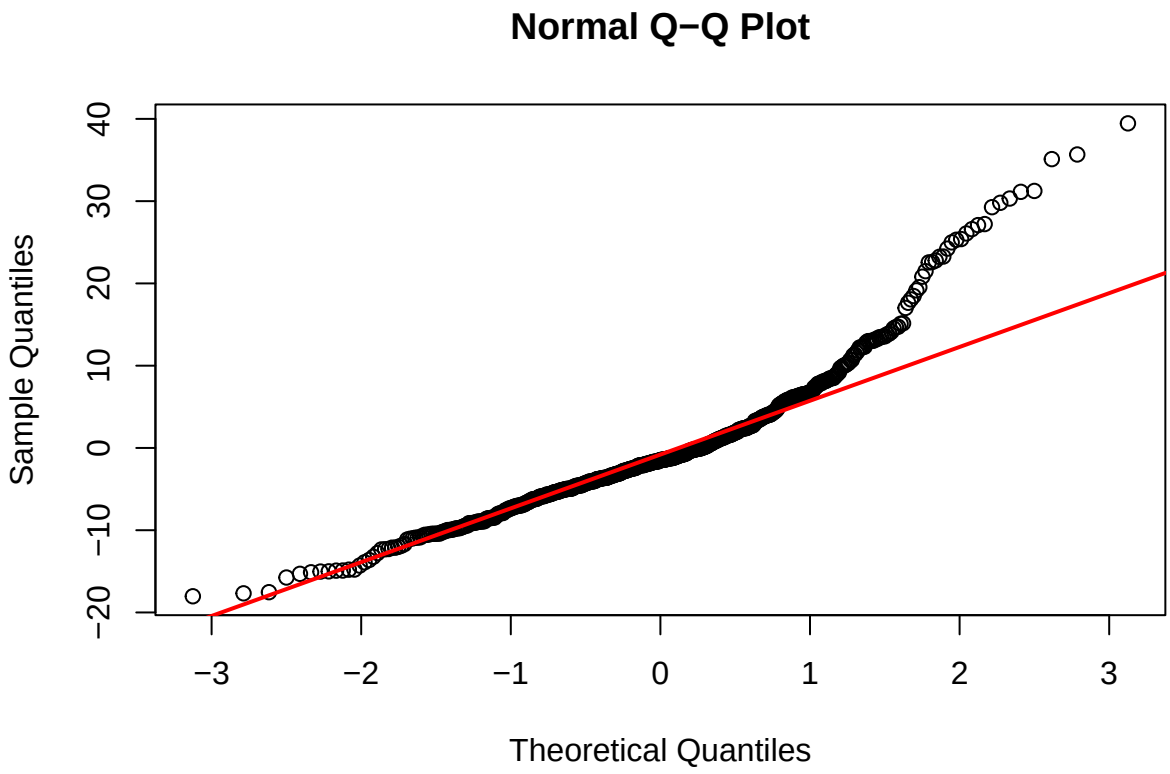



```
print(summary(results2$model))
```

```
##
## Call:
## lm(formula = Age ~ ., data = chosen_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.030  -5.217  -1.486   3.606  39.454
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     32.8342     2.2629  14.510  < 2e-16 ***
## `Primary streaming service`      2.0411     0.3210   6.358 4.30e-10 ***
## Instrumentalist                 -3.7797     0.9112  -4.148 3.88e-05 ***
## `Frequency [Folk]`               1.6907     0.3756   4.501 8.24e-06 ***
## `Frequency [Pop]`               -2.4410     0.4508  -5.415 9.15e-08 ***
## `Hours per day`                 -0.3943     0.1547  -2.549  0.01107 *
## `Frequency [Classical]`          1.5921     0.3984   3.996 7.31e-05 ***
## `Frequency [R&B]`                1.0671     0.3859   2.765  0.00588 **
## `Frequency [Video game music]`  -1.6615     0.3635  -4.570 6.02e-06 ***
## Exploratory                     -1.1332     0.9198  -1.232  0.21847
## Composer                        -0.8382     1.1212  -0.748  0.45506
## OCD                             -0.2546     0.1330  -1.914  0.05608 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.842 on 551 degrees of freedom
## Multiple R-squared:  0.2462, Adjusted R-squared:  0.2312
## F-statistic: 16.36 on 11 and 551 DF,  p-value: < 2.2e-16
```

```
qqnorm(results2$model$residuals)
qqline(results2$model$residuals, col = "red", lwd = 2)
```



Vidimo da je R^2 vrijednost povećana, a reziduali su bliži normalnoj razdiobi. Dodavanjem novih varijabli, prediktivna moć modela je poboljšana, ali je model ipak značajno kompliciraniji. Ukoliko bi nastavili i koristili **SVE** varijable, model bi bio imao najveću prediktivnu moć koju može postići, ali bi bio i najkompleksniji. Naime tim pristupom riskiramo da uzmemo varijable koje su međusobno previše korelirane i time bi narušile interpretaciju modela. U našem slučaju varijable koje smo odabrali nisu značajno korelirane što se može provjeriti korelacijskom matricom na početku ovog zadatka.

Zaključak: Predviđanje dobi putem slušanja glazbe je moguće, ali s ograničenom preciznošću ($R^2 = 0.24$).

Najznačajniji prediktori u višestrukoj regresiji su slijedeći:

- Primary streaming servis (+2.04 godina)
- Frekvencija slušanja popa (-2.44 godina)
- Status instrumentalista (-3.78 godina)
- Frekvencija slušanja folka (+ 1.69 godina)

Model objašnjava oko 24% varijance u dobi, što ukazuje da glazbene preferencije mogu djelomično, ali ne potpuno, predvidjeti dob slušatelja.