Stats 199: Human Insights Through Survey

Sean Nagler

4/16/2020

Abstract:

In 2013, students of a Statistics class at a college university in Slovakia were asked to participate in a survey and invite their friends to do the same. A total of 1,010 people were surveyed and were each asked to answer 150 questions. The questions were broken up by topics, which included music & movie preferences, hobbies, phobias, health habits, personality traits and spending habits. The participants' answers to these questions were then saved and compiled into a dataset that is available at kaggle.com/miroslavsabo/young-people-survey. The goal of this report is to access and analyze this dataset, with a focus on creating and providing graphics & visualations that provide insights into the human condition. While these insights will only apply to this small subset of the human population and cannot represent the human population as whole, they are still useful insights.

Part 1:

Data Acquisition & Cleaning:

```
# Read dataset from Kaggle into R
rm(list = ls())
responses <- read.csv("responses.csv", header = T)
data <- data.frame(responses)

# Show some attributes of data
head(data[1:10], n = 5)
```

```
##   Music Slow.songs.or.fast.songs Dance Folk Country Classical.music Musical Pop
## 1     5                       3     2    1       2               2       1   5
## 2     4                       4     2    1       1               1       2   3
## 3     5                       5     2    2       3               4       5   3
## 4     5                       3     2    1       1               1       1   2
## 5     5                       3     4    3       2               4       3   5
##   Rock Metal.or.Hardrock
## 1    5                 1
## 2    5                 4
## 3    5                 3
## 4    2                 1
## 5    3                 1
```

```
dim(data)
```

```
## [1] 1010  150
```

```r
colnames(data)
```

```
##   [1] "Music"                          "Slow.songs.or.fast.songs"
##   [3] "Dance"                          "Folk"
##   [5] "Country"                        "Classical.music"
##   [7] "Musical"                        "Pop"
##   [9] "Rock"                           "Metal.or.Hardrock"
##  [11] "Punk"                           "Hiphop..Rap"
##  [13] "Reggae..Ska"                    "Swing..Jazz"
##  [15] "Rock.n.roll"                    "Alternative"
##  [17] "Latino"                         "Techno..Trance"
##  [19] "Opera"                          "Movies"
##  [21] "Horror"                         "Thriller"
##  [23] "Comedy"                         "Romantic"
##  [25] "Sci.fi"                         "War"
##  [27] "Fantasy.Fairy.tales"            "Animated"
##  [29] "Documentary"                    "Western"
##  [31] "Action"                         "History"
##  [33] "Psychology"                     "Politics"
##  [35] "Mathematics"                    "Physics"
##  [37] "Internet"                       "PC"
##  [39] "Economy.Management"             "Biology"
##  [41] "Chemistry"                      "Reading"
##  [43] "Geography"                      "Foreign.languages"
##  [45] "Medicine"                       "Law"
##  [47] "Cars"                           "Art.exhibitions"
##  [49] "Religion"                       "Countryside..outdoors"
##  [51] "Dancing"                        "Musical.instruments"
##  [53] "Writing"                        "Passive.sport"
##  [55] "Active.sport"                   "Gardening"
##  [57] "Celebrities"                    "Shopping"
##  [59] "Science.and.technology"         "Theatre"
##  [61] "Fun.with.friends"               "Adrenaline.sports"
##  [63] "Pets"                           "Flying"
##  [65] "Storm"                          "Darkness"
##  [67] "Heights"                        "Spiders"
##  [69] "Snakes"                         "Rats"
##  [71] "Ageing"                         "Dangerous.dogs"
##  [73] "Fear.of.public.speaking"        "Smoking"
##  [75] "Alcohol"                        "Healthy.eating"
##  [77] "Daily.events"                   "Prioritising.workload"
##  [79] "Writing.notes"                  "Workaholism"
##  [81] "Thinking.ahead"                 "Final.judgement"
##  [83] "Reliability"                    "Keeping.promises"
##  [85] "Loss.of.interest"               "Friends.versus.money"
##  [87] "Funniness"                      "Fake"
##  [89] "Criminal.damage"                "Decision.making"
##  [91] "Elections"                      "Self.criticism"
##  [93] "Judgment.calls"                 "Hypochondria"
##  [95] "Empathy"                        "Eating.to.survive"
##  [97] "Giving"                         "Compassion.to.animals"
##  [99] "Borrowed.stuff"                 "Loneliness"
## [101] "Cheating.in.school"             "Health"
## [103] "Changing.the.past"              "God"
```

```
## [105] "Dreams"                          "Charity"
## [107] "Number.of.friends"               "Punctuality"
## [109] "Lying"                           "Waiting"
## [111] "New.environment"                 "Mood.swings"
## [113] "Appearence.and.gestures"         "Socializing"
## [115] "Achievements"                    "Responding.to.a.serious.letter"
## [117] "Children"                        "Assertiveness"
## [119] "Getting.angry"                   "Knowing.the.right.people"
## [121] "Public.speaking"                 "Unpopularity"
## [123] "Life.struggles"                  "Happiness.in.life"
## [125] "Energy.levels"                   "Small...big.dogs"
## [127] "Personality"                     "Finding.lost.valuables"
## [129] "Getting.up"                      "Interests.or.hobbies"
## [131] "Parents..advice"                 "Questionnaires.or.polls"
## [133] "Internet.usage"                  "Finances"
## [135] "Shopping.centres"                "Branded.clothing"
## [137] "Entertainment.spending"          "Spending.on.looks"
## [139] "Spending.on.gadgets"             "Spending.on.healthy.eating"
## [141] "Age"                             "Height"
## [143] "Weight"                          "Number.of.siblings"
## [145] "Gender"                          "Left...right.handed"
## [147] "Education"                       "Only.child"
## [149] "Village...town"                  "House...block.of.flats"
```

The dataset has dimensions 1010 X 150, with 150 questions on the survey and 1010 people having participated.

Looking at the data, the first thing I notice is that almost all the columns are of class integer, meaning the responses are stored on a number scale. For example, one of the questions on the survey is listed as "I enjoy listening to music". The survey subject then input a number from 1 to 5, a "1" meaning they strongly disagree with the statement (really don't enjoy listening to music) and a "5" meaning they really enjoy listening to music. All of the variables are stored in this way, except for the ones that required more than a single number to answer, such as a "Drinking" question on the survey which required an answer in the form of a phrase like "Social drinker". In order to move on, I will first recode these variables as numerical. These variables are "Smoking", "Alchohol", "Punctuality", "Lying", "Internet.usage", "Gender", "Left... right.handed", "Education", "Only.child", "Village... town", and "House... block.of.flats".

```
# Load dplyr package to make this recoding a bit easier
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Recode factor variables
data$Smoking <- recode(data$Smoking,"tried smoking"='2',"current smoker"='4',
                       "former smoker"='3',"never smoked"='1')

data$Alcohol <- recode(data$Alcohol,"drink a lot"='3',"social drinker"='2',"never"='1')

data$Punctuality <- recode(data$Punctuality,"i am often early"='1',
```

```
                            "i am always on time"='2',
                            "i am often running late"='3')

data$Lying <- recode(data$Lying,"everytime it suits me"='4',
                     "only to avoid hurting someone"='3',
                     "sometimes"='2',"never"='1')

data$Internet.usage <- recode(data$Internet.usage,"few hours a day"='3',
                              "less than an hour a day"='2',
                              "most of the day"='4',"no time at all"='1')

data$Gender <- recode(data$Gender,"female"='1',"male"='2')

data$Left...right.handed <- recode(data$Left...right.handed,"left handed"='1',
                                   "right handed"='2')


data$Education <- recode(data$Education, "currently a primary school pupil"='1',
                         "primary school"='2', "secondary school"='3',
                         "college/bachelor degree"='4',"masters degree"='5',
                         "doctorate degree"='6')

data$Only.child <- recode(data$Only.child,"yes"='1',"no"='2')

data$Village...town <- recode(data$Village...town,"village"='1',"city"='2')

data$House...block.of.flats <- recode(data$House...block.of.flats,
                                      "block of flats"='1',"house/bungalow"='2')
```

Now, all the observations that had worded responses are now seen as numerical responses to match the rest of the dataset.

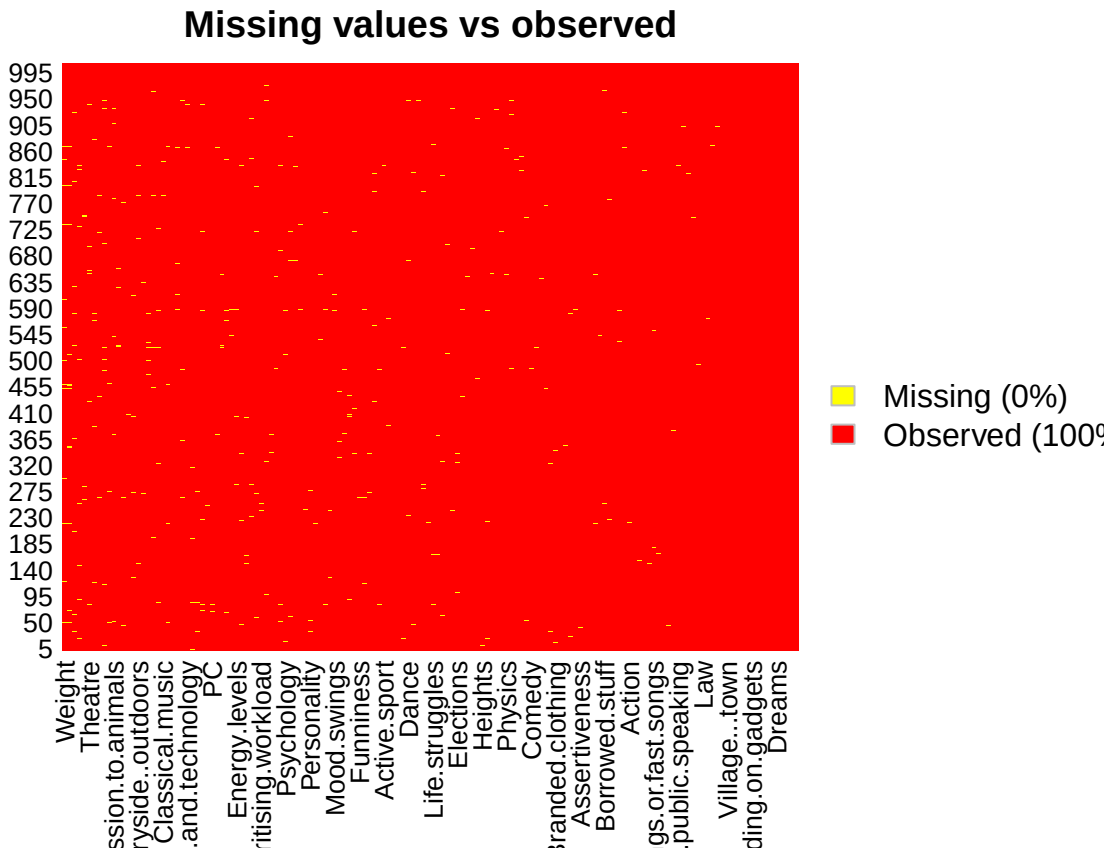
My next step in data cleaning is to check for NAs. I will do this using the Amelia package and missmap() function.

```
# Create missing value map
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(data, main = "Missing values vs observed", col = c("yellow", "red"))
```



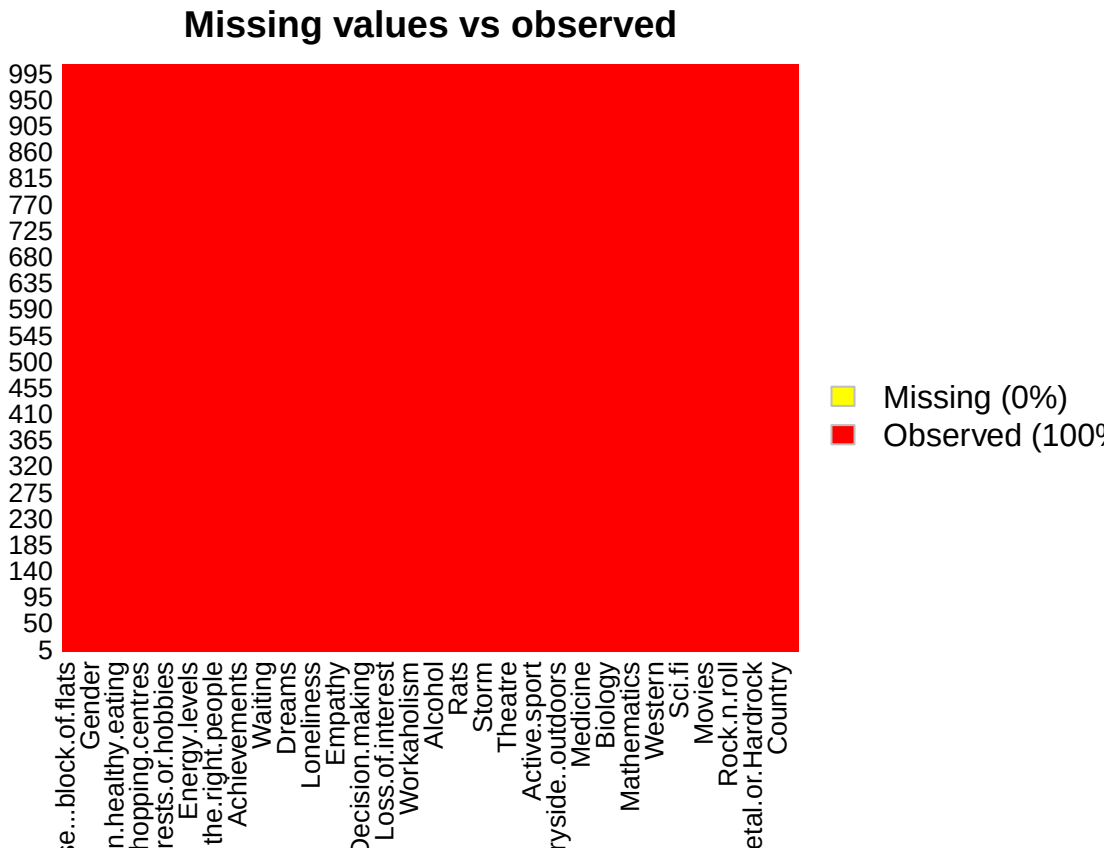
From looking at what equates to this "missing value map", we can see that there are missing values or NAs scattered throughout the entire dataset (shown by the yellow spaces in the visual) . Instead of removing them all and creating an issue of unequal column lengths as well altering any future findings, I will instead perform median imputation for each variable. This will replace each NA with the median of the column that particular NA is in. In order to do this though, I will need every column to be of class integer, so I first need to change the class of the variables that I recoded just before into class integer.

```
# Convert variables stored as characters into integers in order to perform median imputation
data$Smoking <- as.integer(as.character(data$Smoking))

data$Alcohol <- as.integer(as.character(data$Alcohol))

data$Punctuality <- as.integer(as.character(data$Punctuality))

data$Lying <- as.integer(as.character(data$Lying))

data$Internet.usage <- as.integer(as.character(data$Internet.usage))

data$Gender <- as.integer(as.character(data$Gender))

data$Left...right.handed <- as.integer(as.character(data$Left...right.handed))

data$Education <- as.integer(as.character(data$Education))

data$Only.child <-  as.integer(as.character(data$Only.child))

data$Village...town <- as.integer(as.character(data$Village...town))
```

```
data$House...block.of.flats <-  as.integer(as.character(data$House...block.of.flats))

# Now, perform median imputation on every variable

# Create function for median imputation
MedianImpute <- function(df){
  a <- apply(df,2,function(x){x[is.na(x)] <- median(as.numeric(x),na.rm=T);x})
  a1 <- data.frame(a)
  return(a1)
}

# Perform median imputation on dataset and create new filled dataset
filled.data <- MedianImpute(data)

# Check to make sure new dataset has same dimensions and no NAs
dim(filled.data)
```

```
## [1] 1010  150
```

```
missmap(filled.data, main = "Missing values vs observed", col = c("yellow", "red"))
```



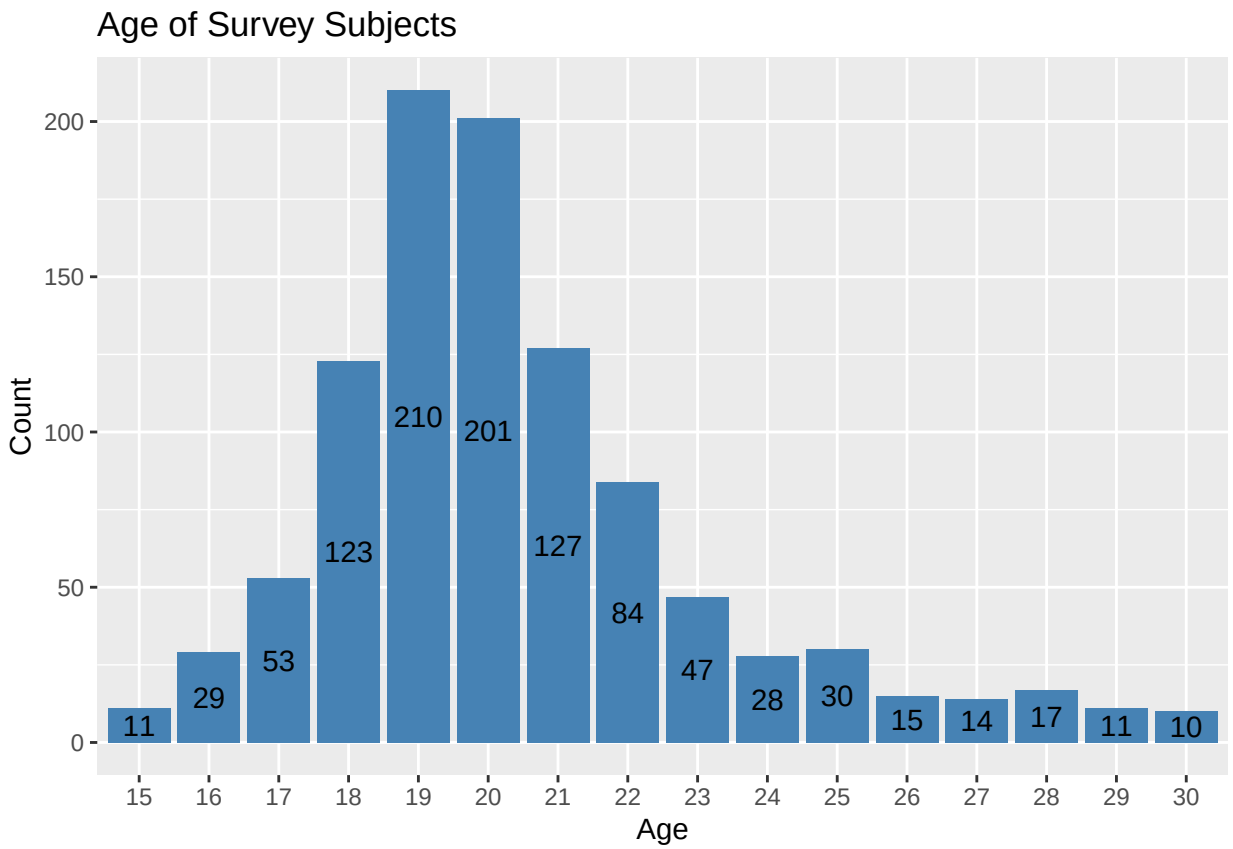
The dimensions remain unchanged and as seen from the updated missing values map, there are no longer missing values (no yellow spaces), so we can move on.

The survey questions were grouped by topic (music, movies, hobbies, etc.) as mentioned before, so I will group the dataset in the same way, creating a dataframe for each survey topic.

```
# Group dataset by survey topics
music.pref <- filled.data[,1:19] # Music preferences
movies.pref <- filled.data[,20:31] # Movie preferences
hobbies <- filled.data[,32:63] # Hobbies & interests
phobias <- filled.data[,64:73] # Phobias
health <- filled.data[,74:76] # Health habits
personality <- filled.data[,77:133] # Personality traits, views on life, & opinions
spending <- filled.data[134:140] # Spending habits
demog <- filled.data[,141:150] # Demographics
```

Before getting into any analysis, it's important to understand where the data is coming from. As mentioned earlier, this survey was given to people in Slovakia so it'll be interesting to see if any patterns found are similar or dissimilar to those known in America. Although the survey didn't ask too much more about identifying demographic factors, it did ask for the participant's age. Knowing which age group these answers come from will help us better understand any conclusions made from the findings.

```
# Create bargraph displaying ages of participants
library(ggplot2)
a <- ggplot(data.frame(filled.data), aes(x=as.factor(demog[,1]))) +
  geom_bar(fill = "steelblue") +
  labs(title = "Age of Survey Subjects", x = "Age",  y = "Count")
a + geom_text(stat = "count", aes(label = ..count.., y = ..count..), position = position_stack(0.5))
```



Looking at this barplot of the ages of the survey participants, it is clear that the majority of the people who partook are between the ages of 18 & 21, with the youngest participants being 15 and the eldest being 30. So, any conclusion made must take into account that it is derived from data about a young population.

Step 2:

Analyses & Results:

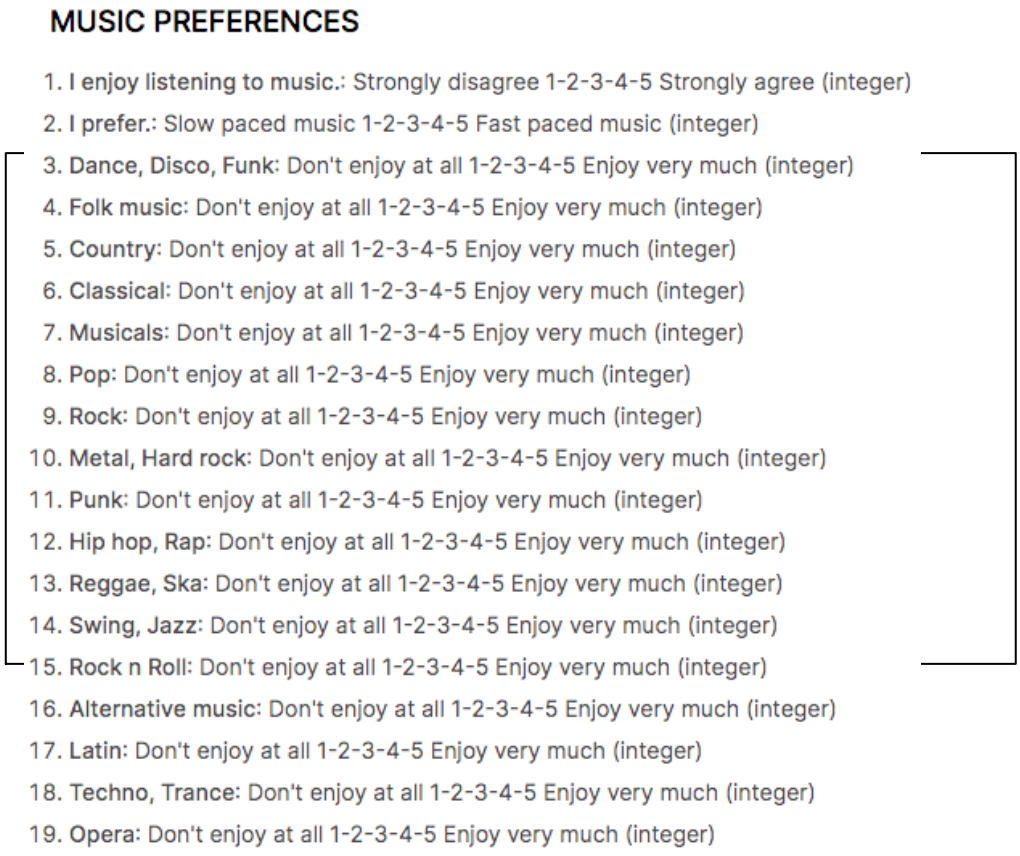
The rest of this report will follow the same structure as the survey, with it being separated by topic. The responses to each topic of the survey will be analyzed and visualized to learn key insights about the aggregation of humans who were surveyed.

Music Preferences:

I'll start by looking at the first area of the survey, which asked its participants about their music preferences.

Seen below is the list and format of the section on the survey called 'Music Preferences'.

```
# Load screenshot of questions for music section
library(png)
img <- readPNG("MusicPref.png")
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n", yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img)
```



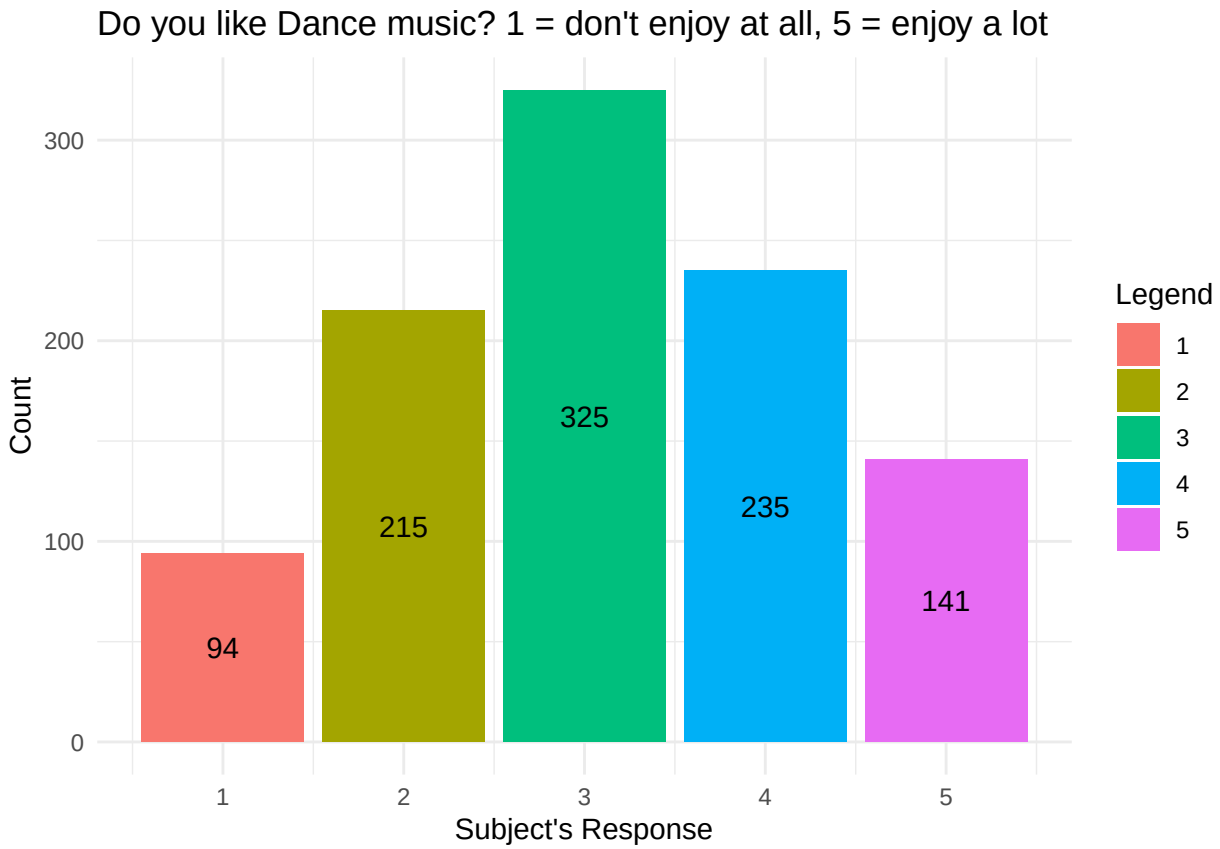
I'll be looking to see if there are any trends in music consumption amongst this subpopulation.

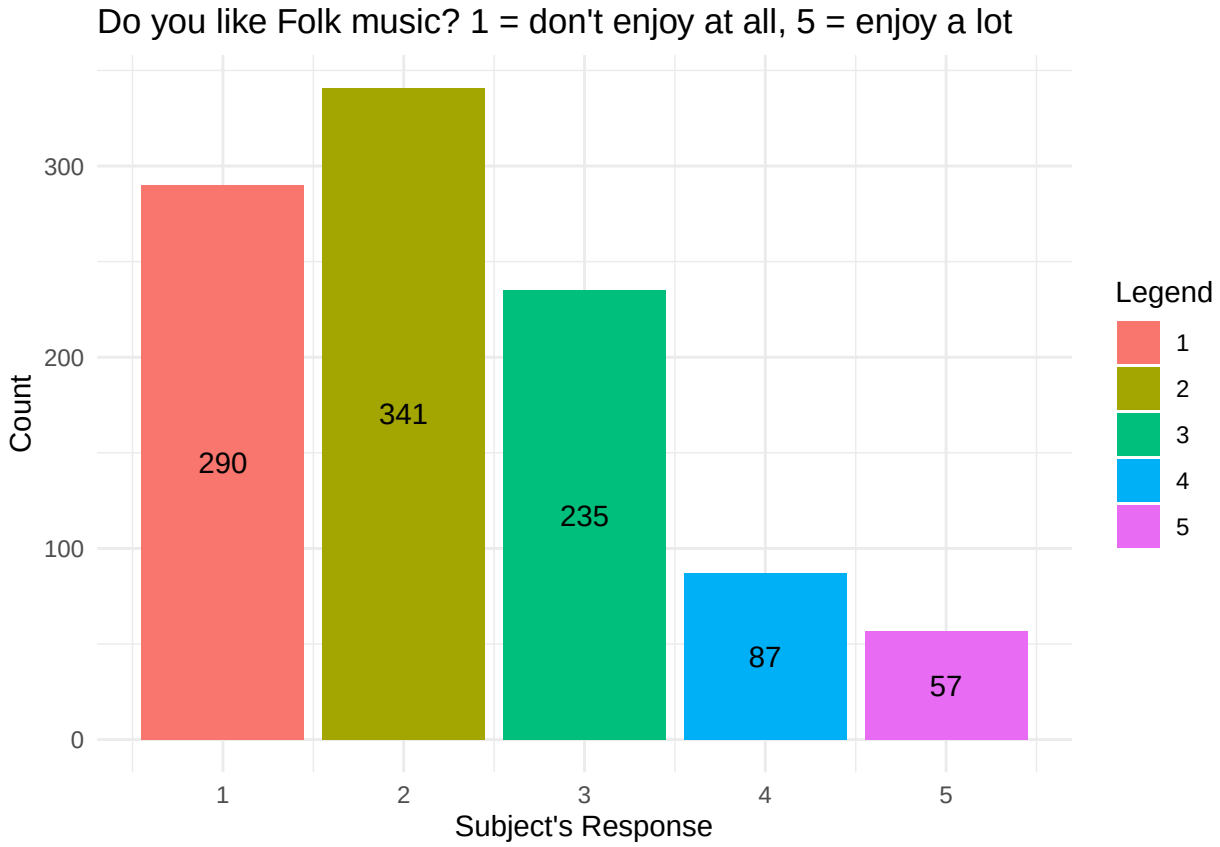
First, let's see which genre of music is most popular among this subpopulation in Slovakia by looking at some bar graphs showing the distribution of responses in this section of the survey.

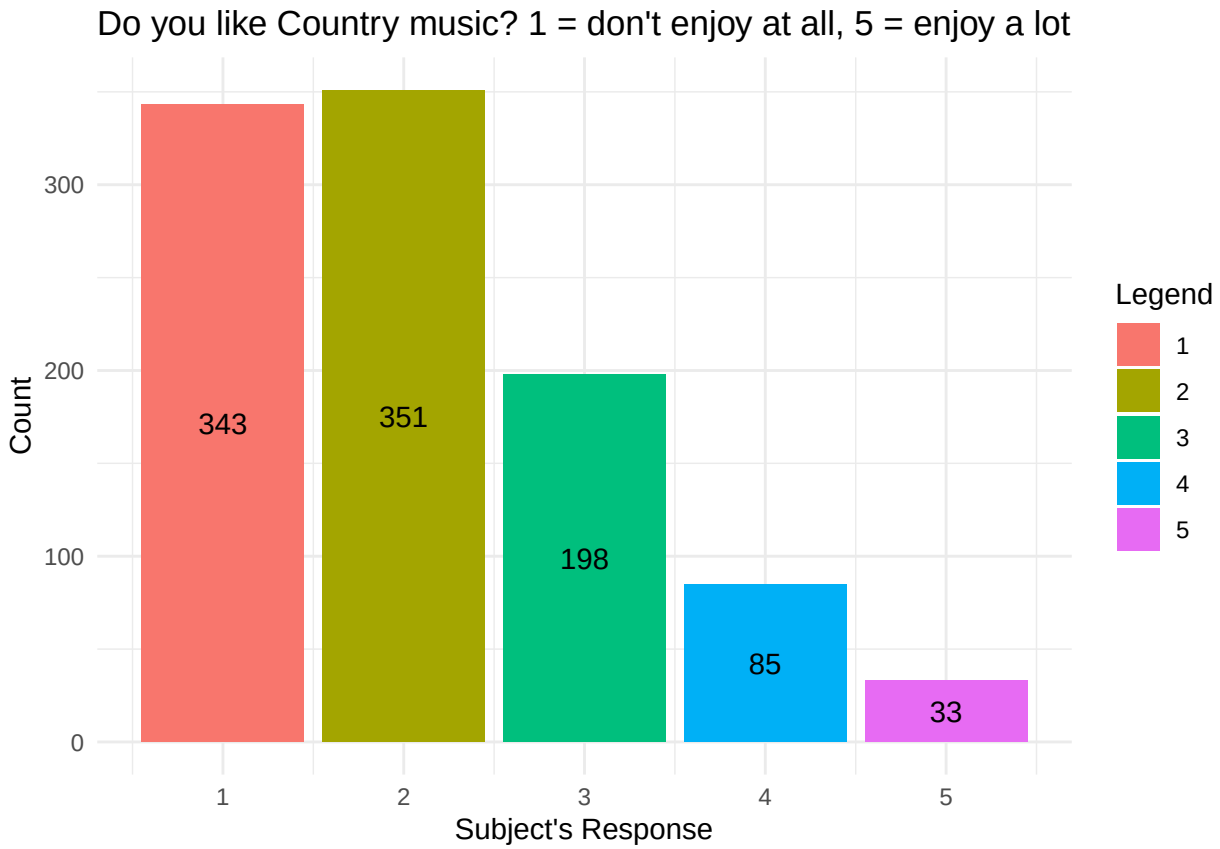
```
library(plyr)
```

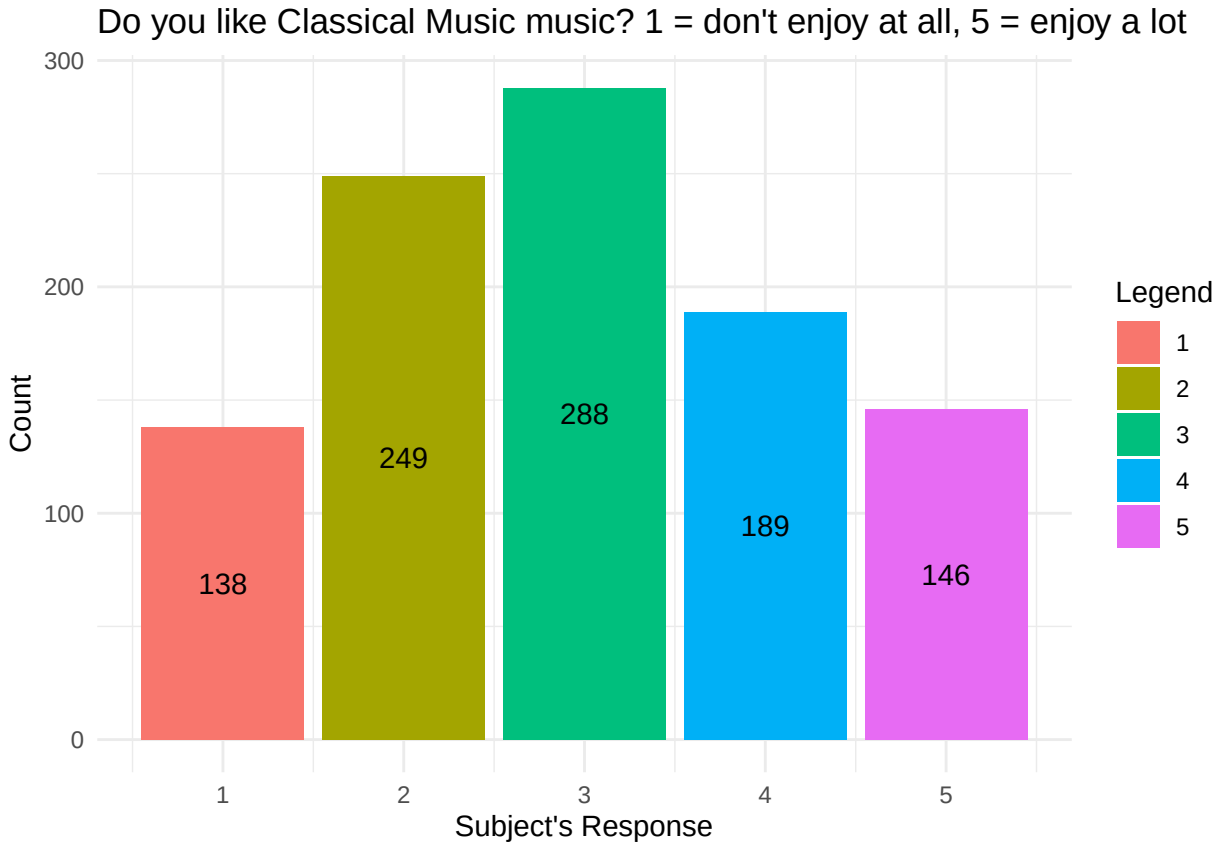
```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

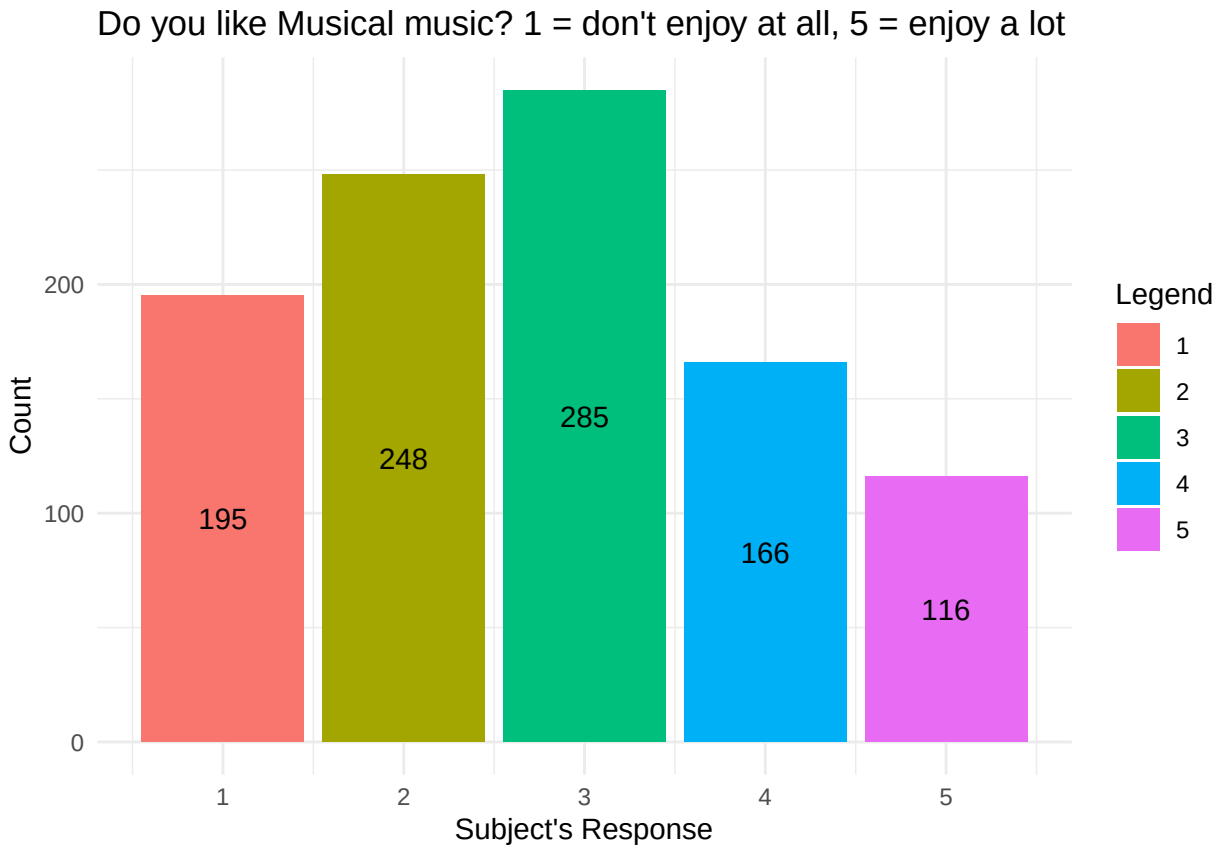
```r
# Create vector of genre names
music.genres <- colnames(music.pref[,3:ncol(music.pref)])

# Rename some of the genres to make it look neater
music.genres = replace(music.genres, music.genres == "Classical.music", "Classical Music")
music.genres = replace(music.genres, music.genres == "Metal.or.Hardrock", "Metal/Hardrock")
music.genres = replace(music.genres, music.genres == "Hiphop..Rap", "Hiphop/Rap")
music.genres = replace(music.genres, music.genres == "Reggae..Ska", "Reggae")
music.genres = replace(music.genres, music.genres == "Swing..Jazz", "Jazz")
music.genres = replace(music.genres, music.genres == "Techno..Trance", "Techno/Trance")
music.genres = replace(music.genres, music.genres == "Rock.n.roll", "Rock & Roll")

# For loop to create bar graphs of responses for each genre
for(i in 3:ncol(music.pref)){

plot <- ggplot(data.frame(music.pref), aes(x = music.pref[,i], fill = factor(music.pref[,i]))) +
  geom_bar() +
  theme_minimal() +
  geom_text(stat = "count", aes(label = ..count.., y = ..count..), position = position_stack(0.5)) +
  labs(title = paste("Do you like", music.genres[i-2], "music? 1 = don't enjoy at all, 5 = enjoy a lot"),
       x = "Subject's Response",  y = "Count", fill = "Legend")

par(mfrow = c(6,3))
print(plot)
}
```

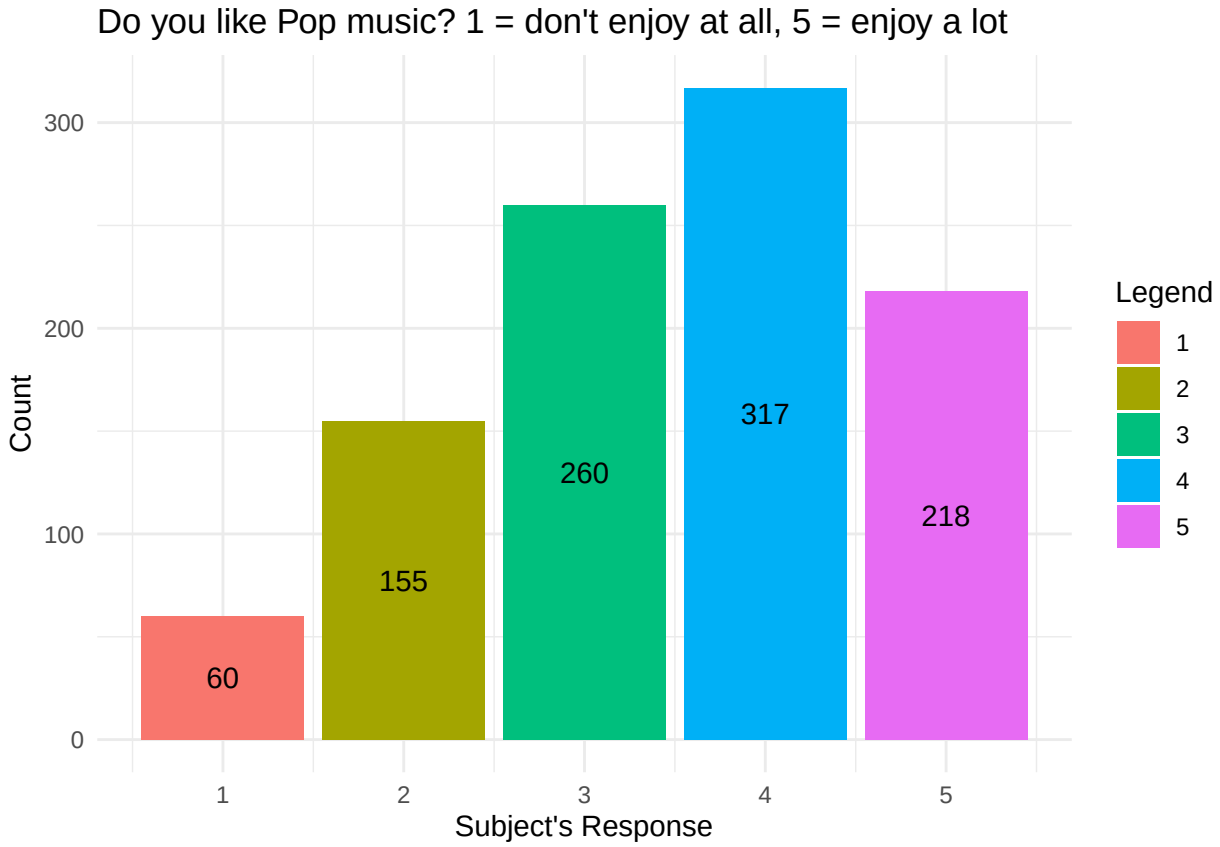

Do you like Dance music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
94
215
325
235
141
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

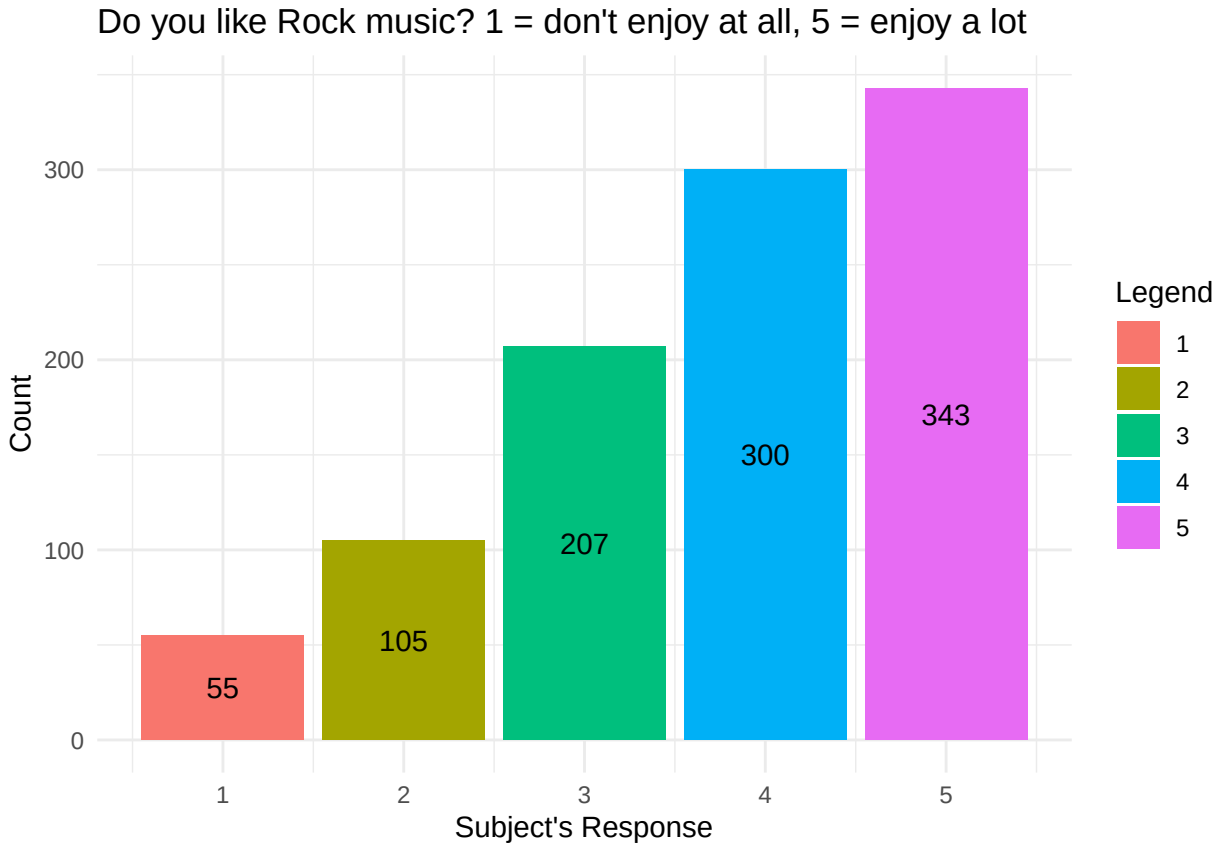

Do you like Folk music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
290
341
235
87
57
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

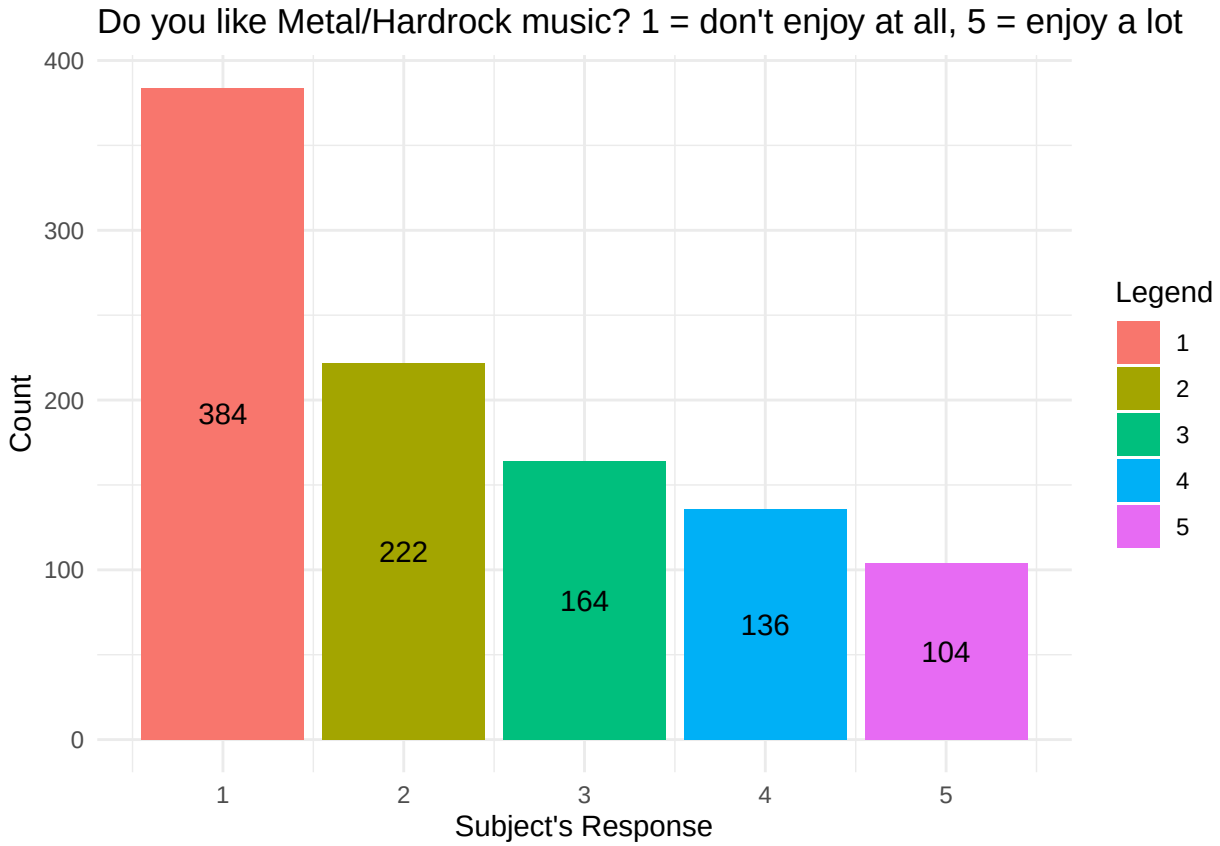

Do you like Country music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
343
351
198
85
33
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

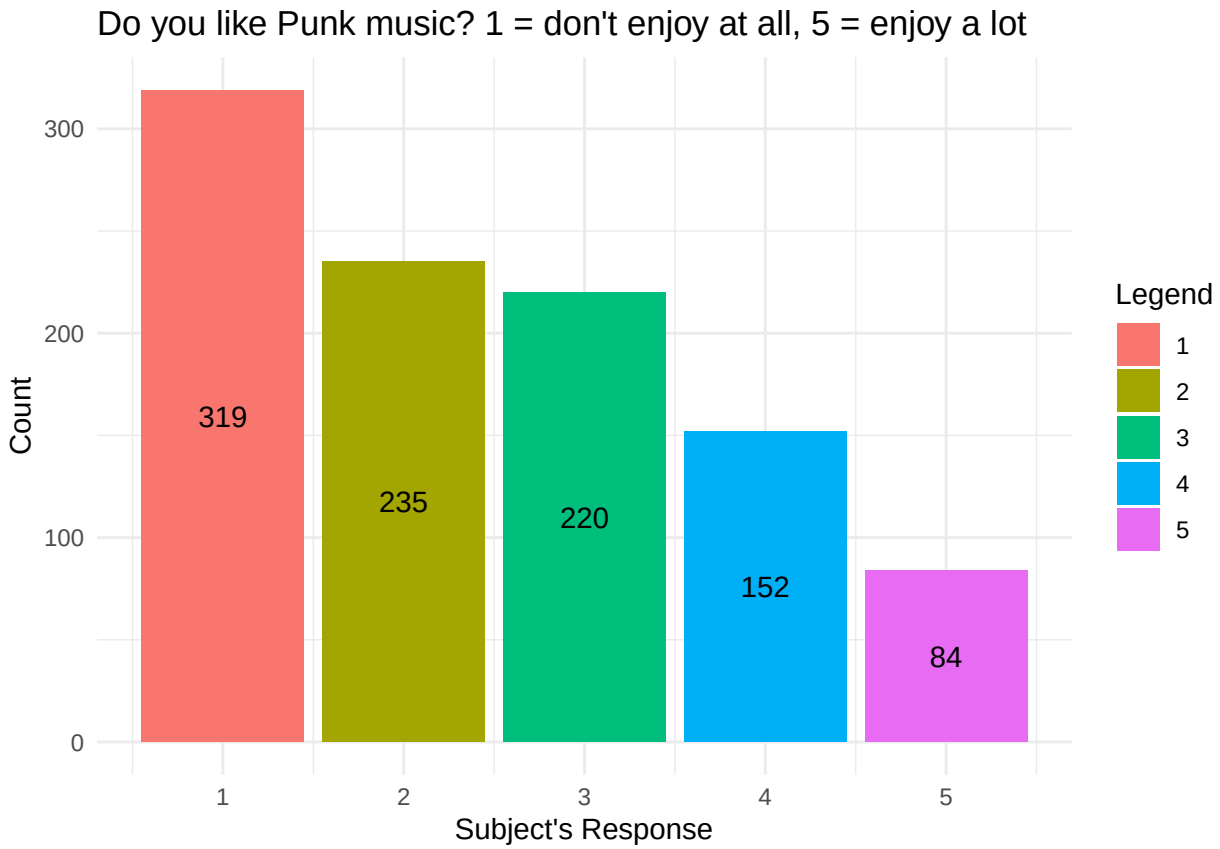

Do you like Classical Music music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
138
249
288
189
146
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

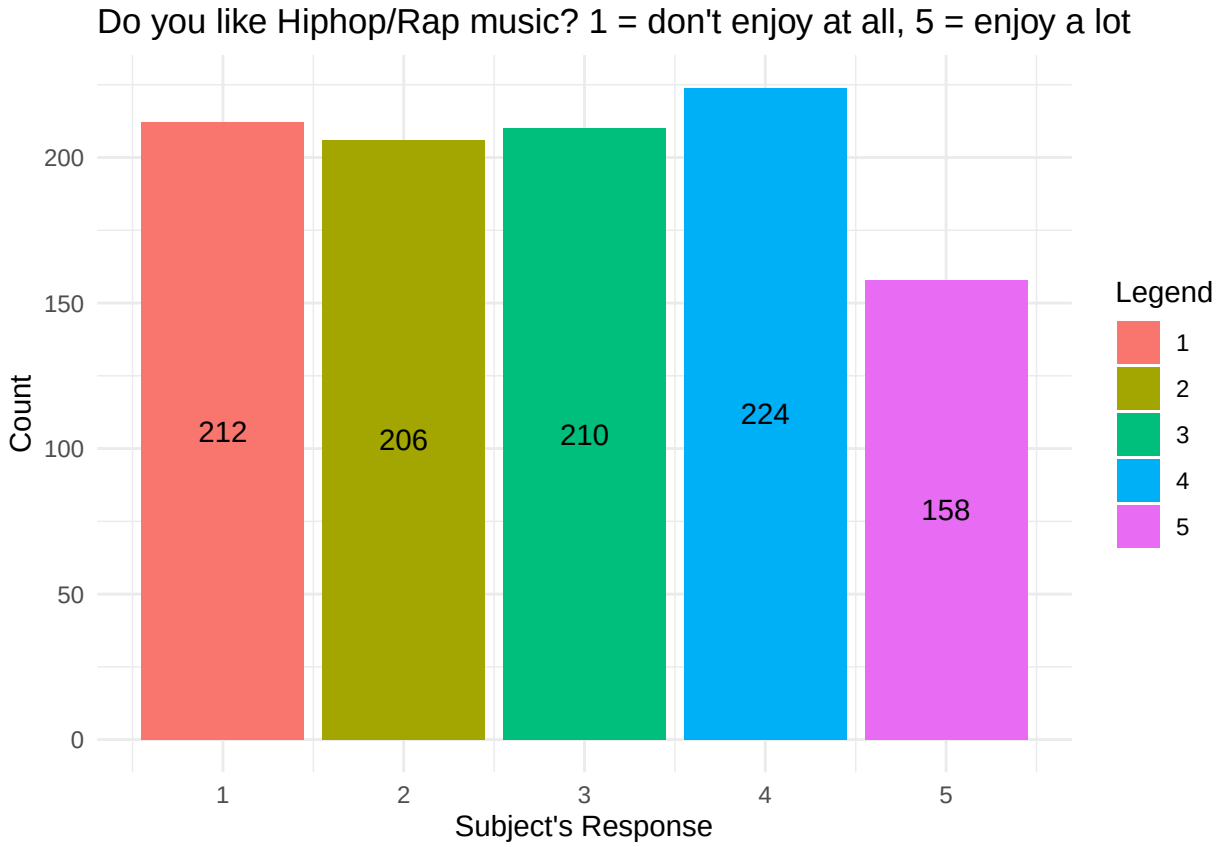

Do you like Musical music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
0
100
200
195
248
285
166
116
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

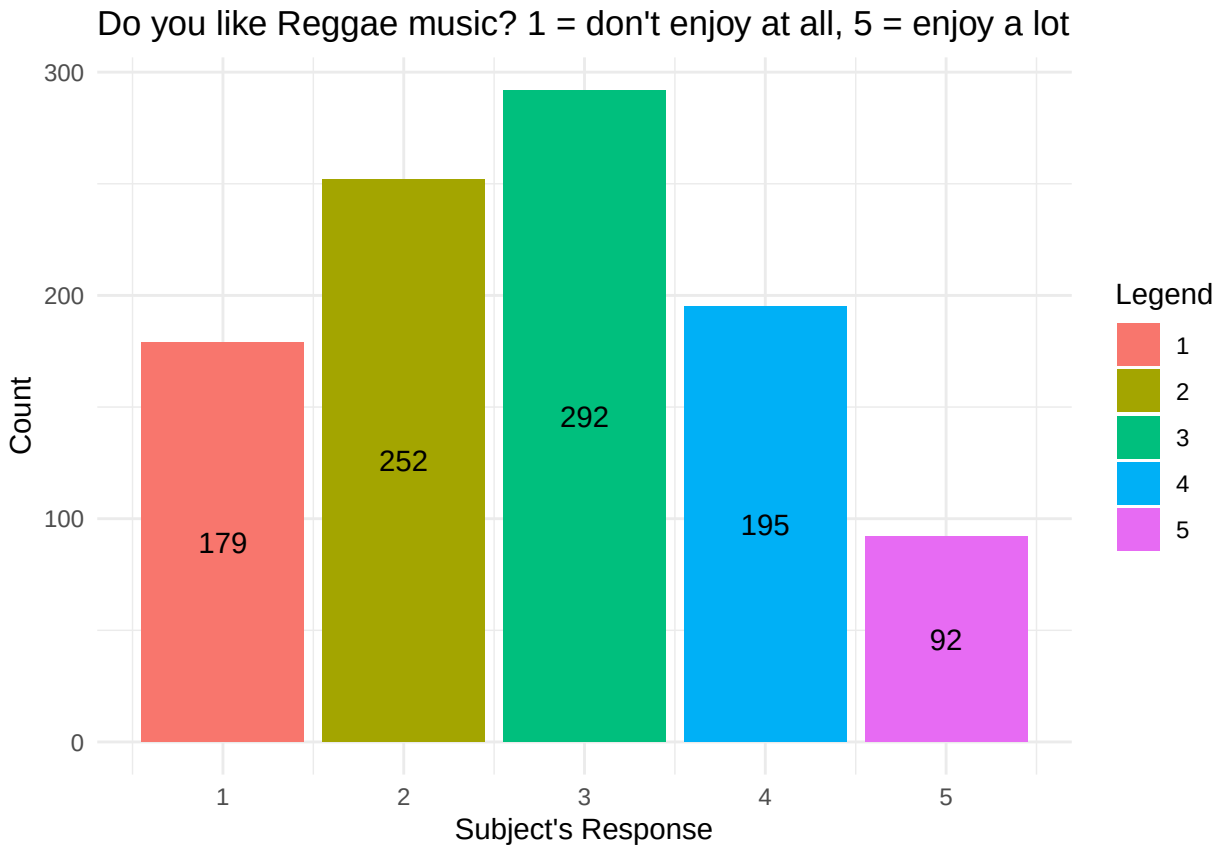

Do you like Pop music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
0
100
200
300
60
155
260
317
218
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

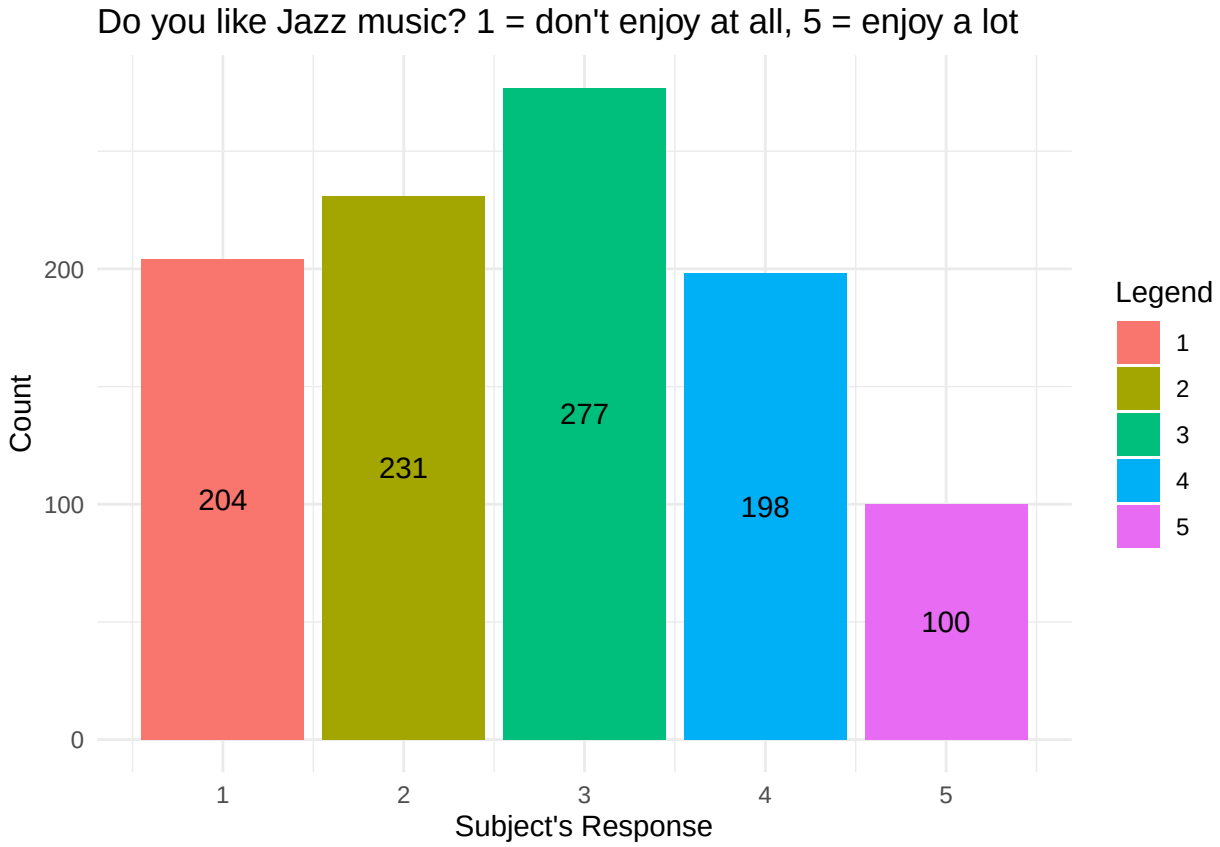

Do you like Rock music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
55
105
207
300
343
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

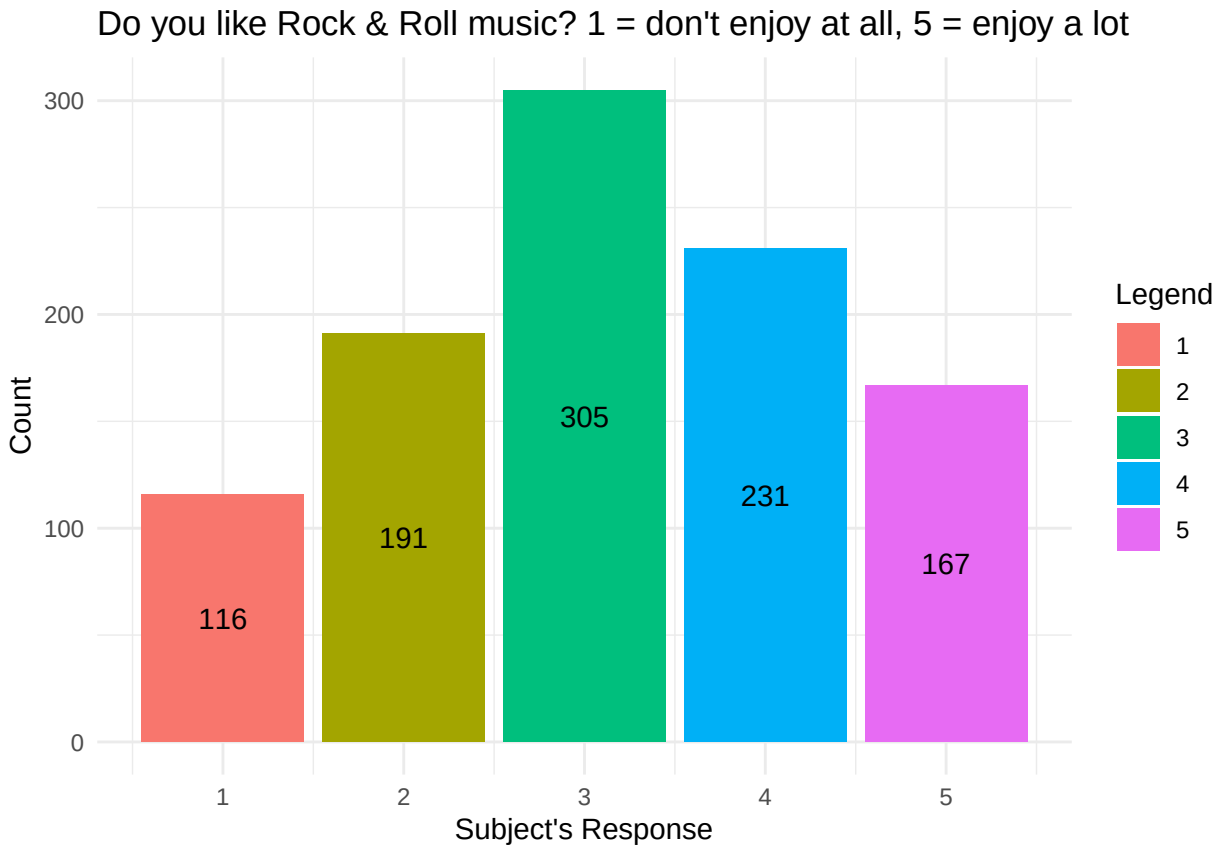

Do you like Metal/Hardrock music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
400
300
200
100
0
384
222
164
136
104
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

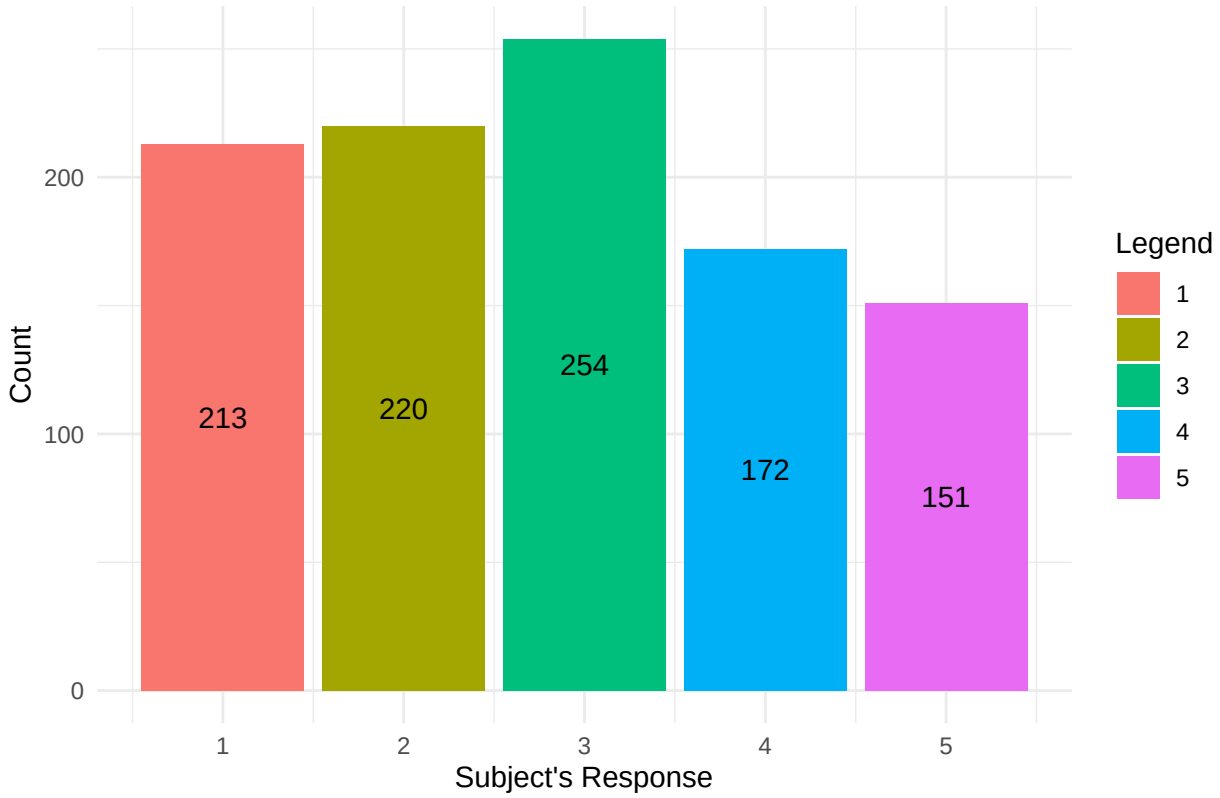

Do you like Punk music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
319
235
220
152
84
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

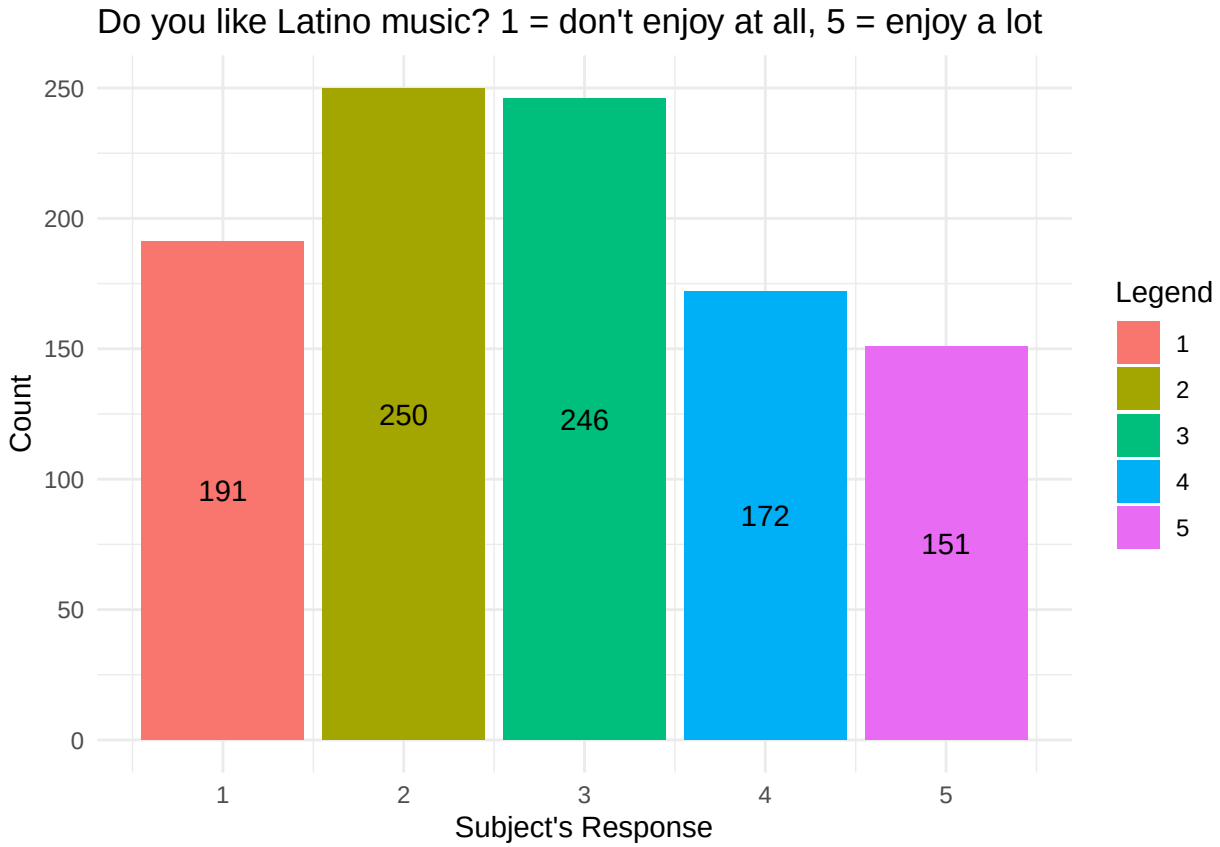

Do you like Hiphop/Rap music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
200
150
100
50
0
212
206
210
224
158
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

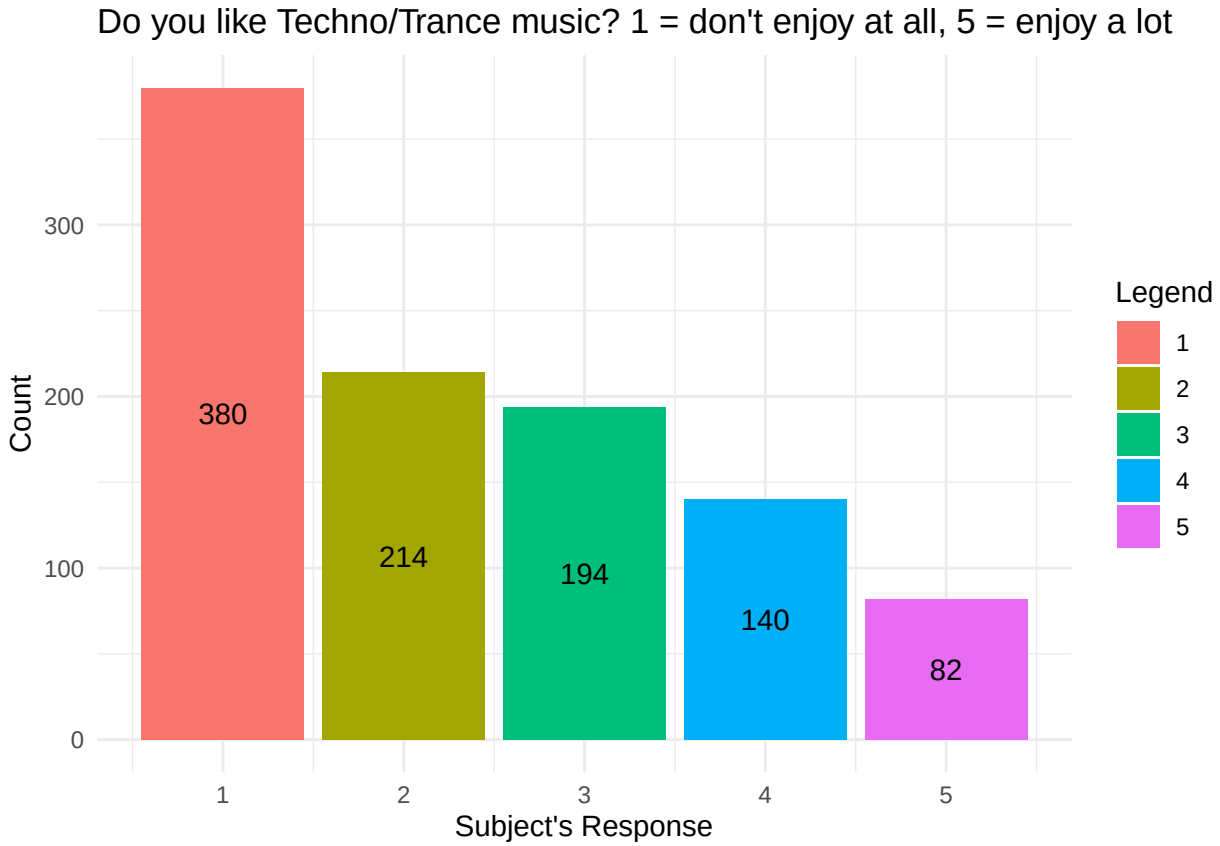

Do you like Reggae music? 1 = don't enjoy at all, 5 = enjoy a lot
300
200
100
0
Count
179
252
292
195
92
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

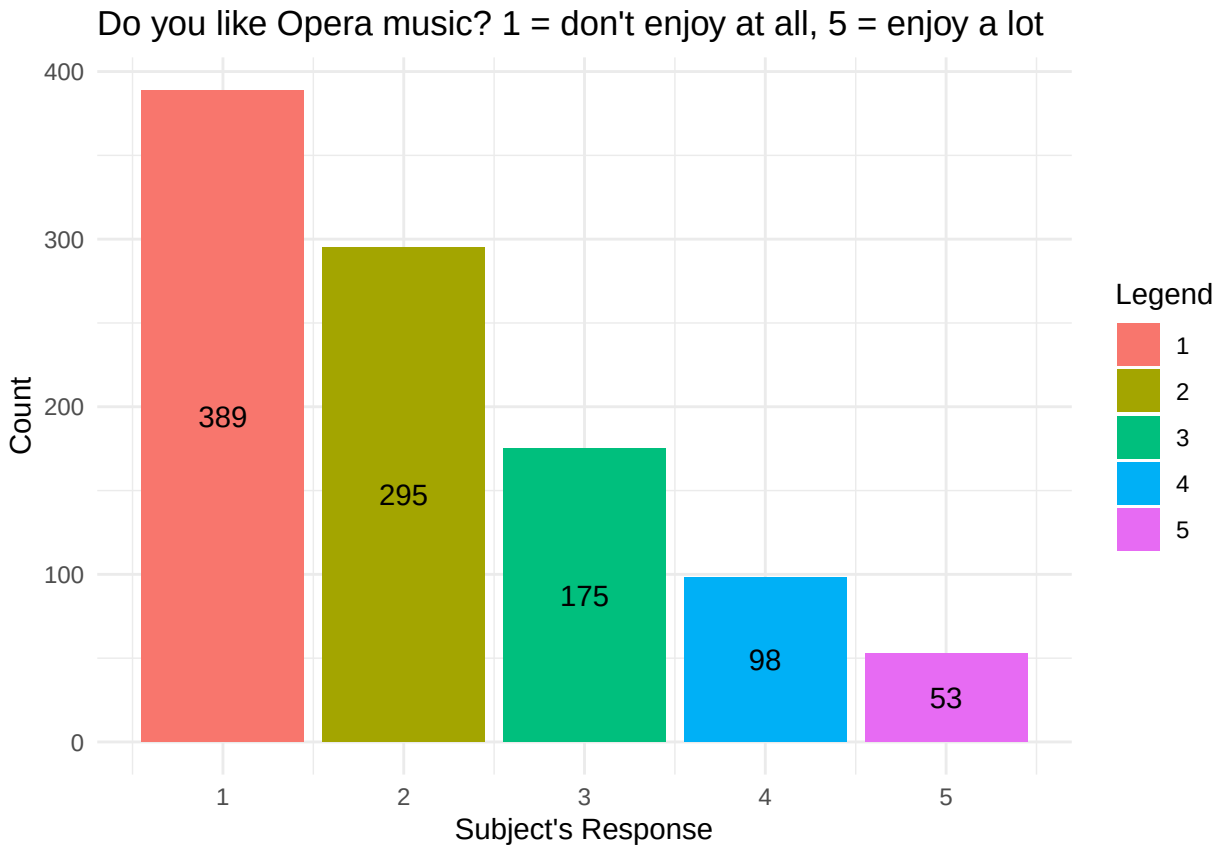

Do you like Jazz music? 1 = don't enjoy at all, 5 = enjoy a lot
200
100
0
Count
204
231
277
198
100
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Rock & Roll music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
116
191
305
231
167
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Alternative music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
200
100
0
213
220
254
172
151
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Latino music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
250
200
150
100
50
0
191
250
246
172
151
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Techno/Trance music? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
380
214
194
140
82
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5



Seen above are bar graphs showing the responses to each genre of music, with a 1 meaning 'don't enjoy it at all' and a 5 meaning 'enjoy it a lot'. Inside each bar is the count or number of participants that answered that question with each respective option. As seen from the graphs:

Techno, Punk, Metal, and Opera music are the least favorites among the participants, with the majority of the responses for each of these genres being a 1.

The overwhelming favorite genre is Rock, with it being the only genre with the majority of reponses being a 5.

Hip hop/Rap music is the most the most evenly distributed bar graph, meaning not a lot of people love it while at the same time, not a lot of people hate it. This is interesting because rap music is very popular in California, but clearly not so much in Slovakia.

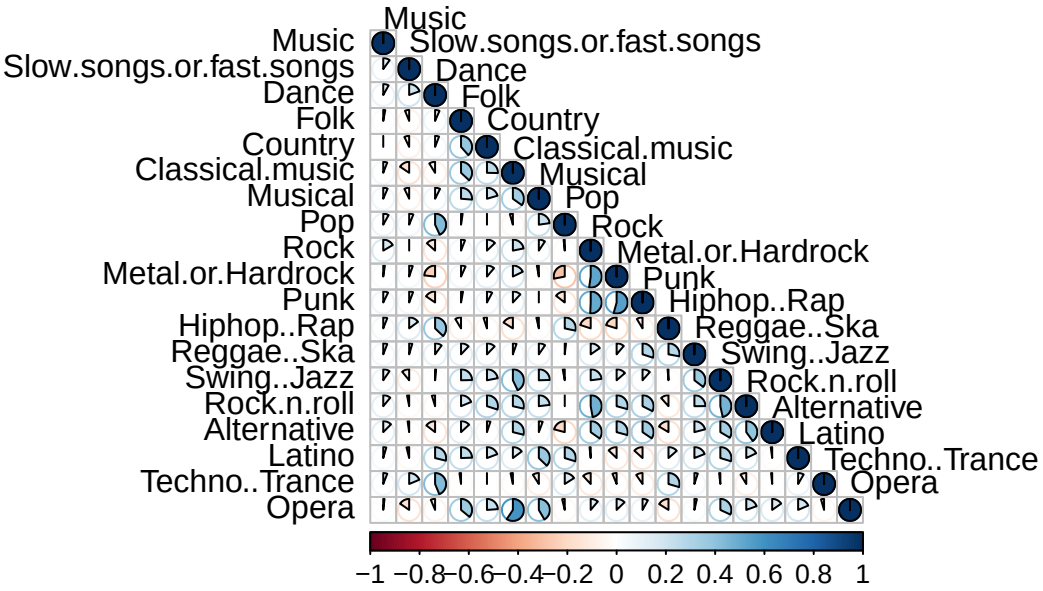
An interesting part about the human connection with music is that each person has a very unique taste in music. No two people's list of top ten songs, for example, are exactly the same. This notion gave me the idea to observe whether there is any connection or pattern between genres. In other words, I'd like to see if liking one genre makes an indiviudal more or less likely to enjoy a different genre.

I will do this using a correlation plot.

```
# Create correlation plot for music preferences
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(music.pref), type = "lower", method = "pi", tl.col = "black", tl.srt = .20,
         title = "Correlation Between Music Preferences", mar=c(0,0,1,0))
```



To provide some clarity on this plot, it is a matrix visualiztion of the correlation matrix between the responses from the music preferences section on the survey. A positive correlation is shown by a blue circle and a negative correlation is shown by an orange circle. The amount of the circle that is filled represents how high this correlation coefficient is. The higher the correlation coefficient, the stronger the relation between the two genres is.

From this, we can see that the highest positive correlation amongst music preferences from this population is between Opera and Classical Music. This shows us that if a person enjoys listening to Opera music, than they will most likely also enjoy listening to Classical Music.

Furthermore, the highest negative correlation is shown to be between Pop and Metal.or.Hardrock, meaning a person who likes Pop music most likely does not like Metal or Hardrock music.

A few more interesting insights from this plot:

Those who enjoy Rock and Roll music also tend to enjoy Jazz music.

Those who like Rap music don't generally like Rock music or Metal or Hardrock music.

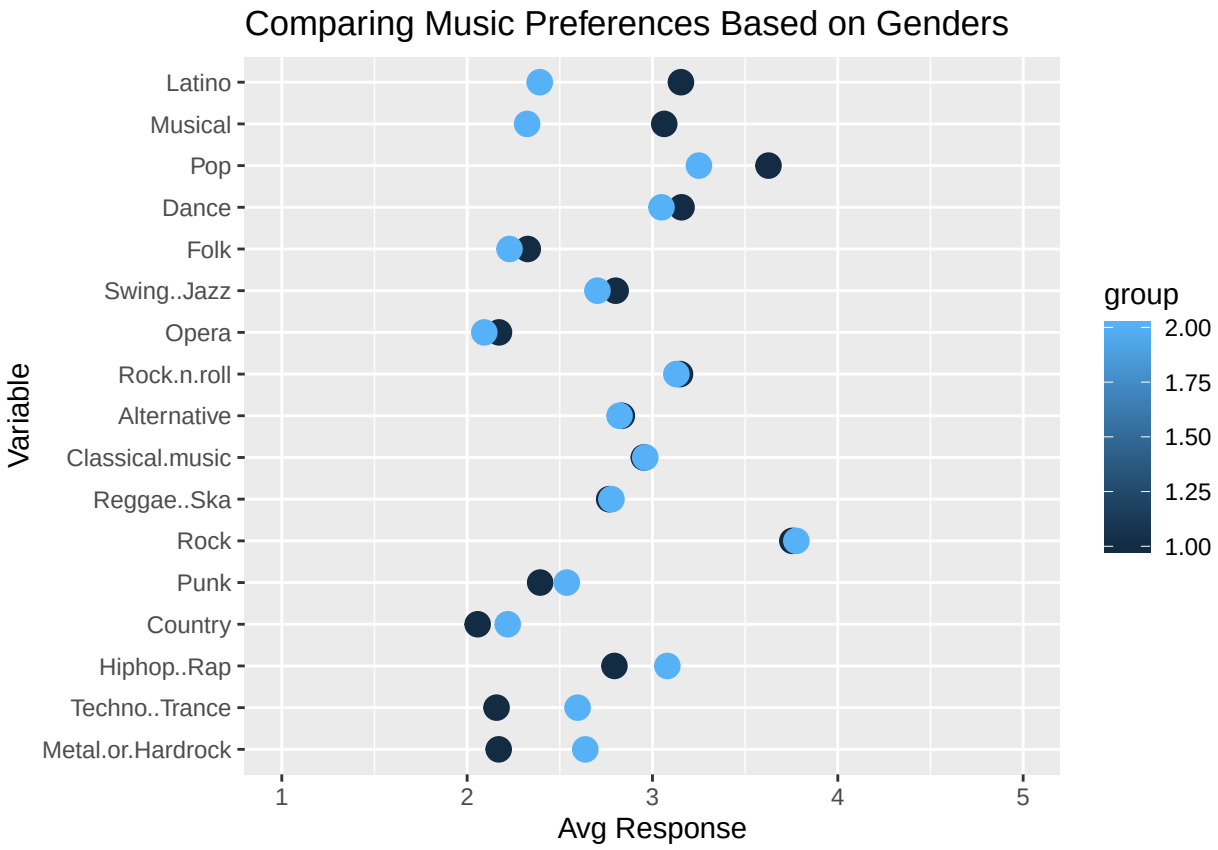
I'd like to break this down a little further. I'd like to see if these preferences depend at all on gender.

```
library(tidyr)

# Create function that splits average responses by gender and plots findings
gender.diff <- function(df, group, start, end){
  averages <- df %>% rename_(group = group) %>%
    select_("group", paste0("`", start, "`:`", end, "`")) %>%
    group_by(group) %>% summarise_all(mean)

  differences <- averages %>% dplyr::select(-group) %>%
    apply(2,function(x) x[1] - x[2]) %>%
    sort %>% names

  averages %>% gather(Variable, `Avg Response`, -group) %>%
```

```
    ggplot(aes(x = Variable, y = `Avg Response`, group = group, colour = group)) +
    geom_point(size = 4) +
    scale_x_discrete(limits = differences) + ylim(1,5) +
    coord_flip()

}

# Run gender differene function on music section of dataset
filled.data %>% gender.diff("Gender", "Dance", "Opera") +
  ggtitle("Comparing Music Preferences Based on Genders")
```

```
## Warning: rename_() is deprecated.
## Please use rename() instead
##
## The 'programming' vignette or the tidyeval book can help you
## to program with rename() : https://tidyeval.tidyverse.org
## This warning is displayed once per session.
```

```
## Warning: select_() is deprecated.
## Please use select() instead
##
## The 'programming' vignette or the tidyeval book can help you
## to program with select() : https://tidyeval.tidyverse.org
## This warning is displayed once per session.
```



As seen earlier, I recoded the Gender variable so that "Female" = 1 and "Male" = 2. So, the average response of females is shown on this plot by the dark blue circles and that of males shown by the light blue circle. So,

looking at this visual, it can be seen that the largest difference in music preferences among sexes is in the Latino and Musical genres, with females preferring both more than males. It can also be seen that males prefer Techno music and Metal music more than females. Another interesting thing to note is that Rock music got very high responses from both male and female participants, confirming what the bargraphs before showed that its the most popular genre amongst this subpopulation.

Movie Preferences:

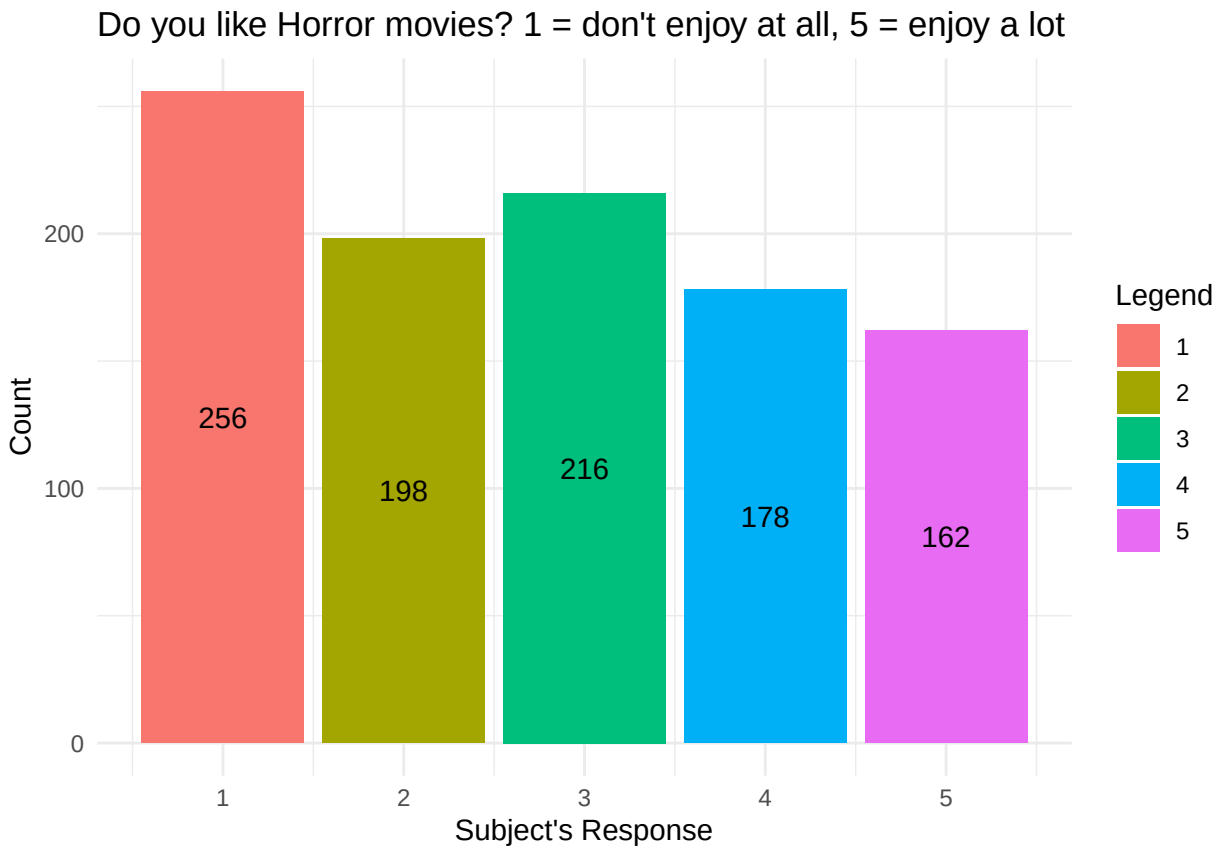
Let's repeat the same process as we did for music, but for movie preferences of this subpopulation, looking for any patterns or insights.

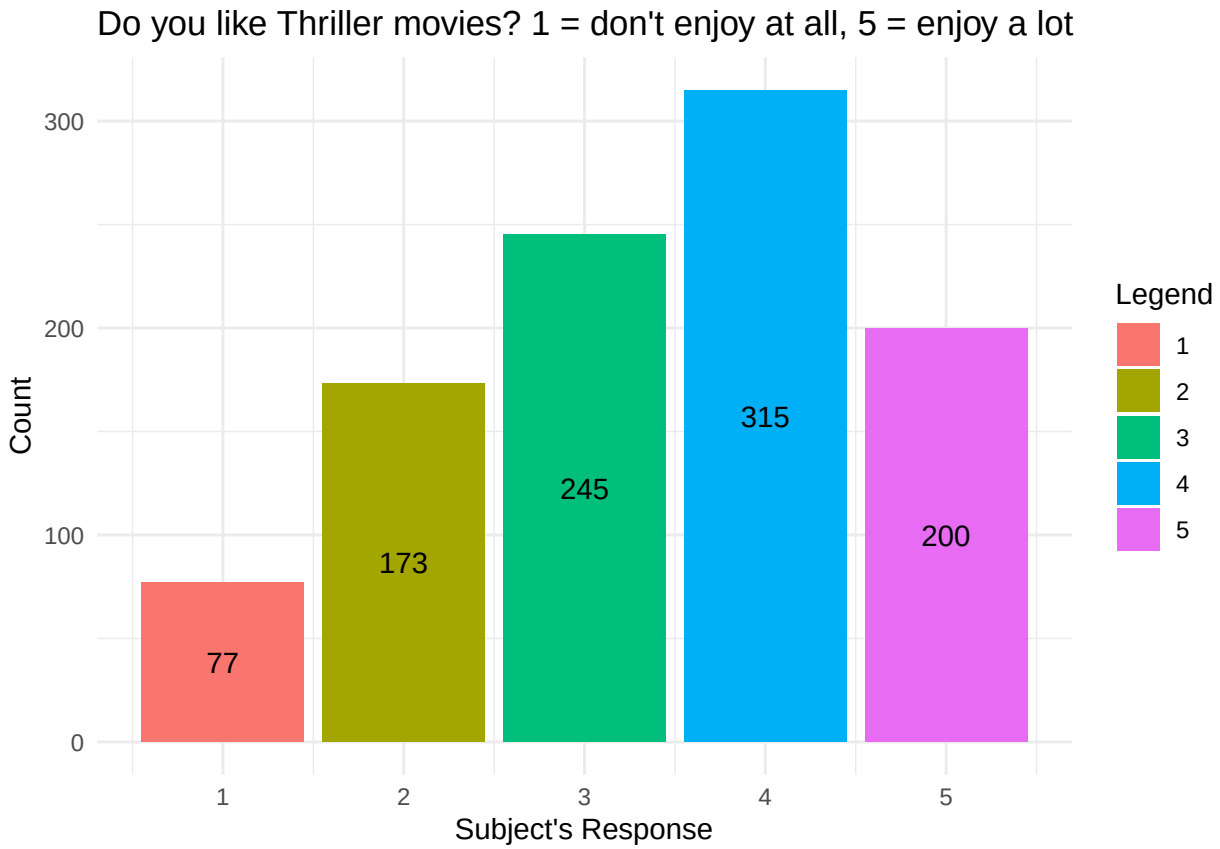
```
# Load screenshot of questions for movie section of survey
library(png)
img <- readPNG("MoviePref.png")
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n", yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img)
```

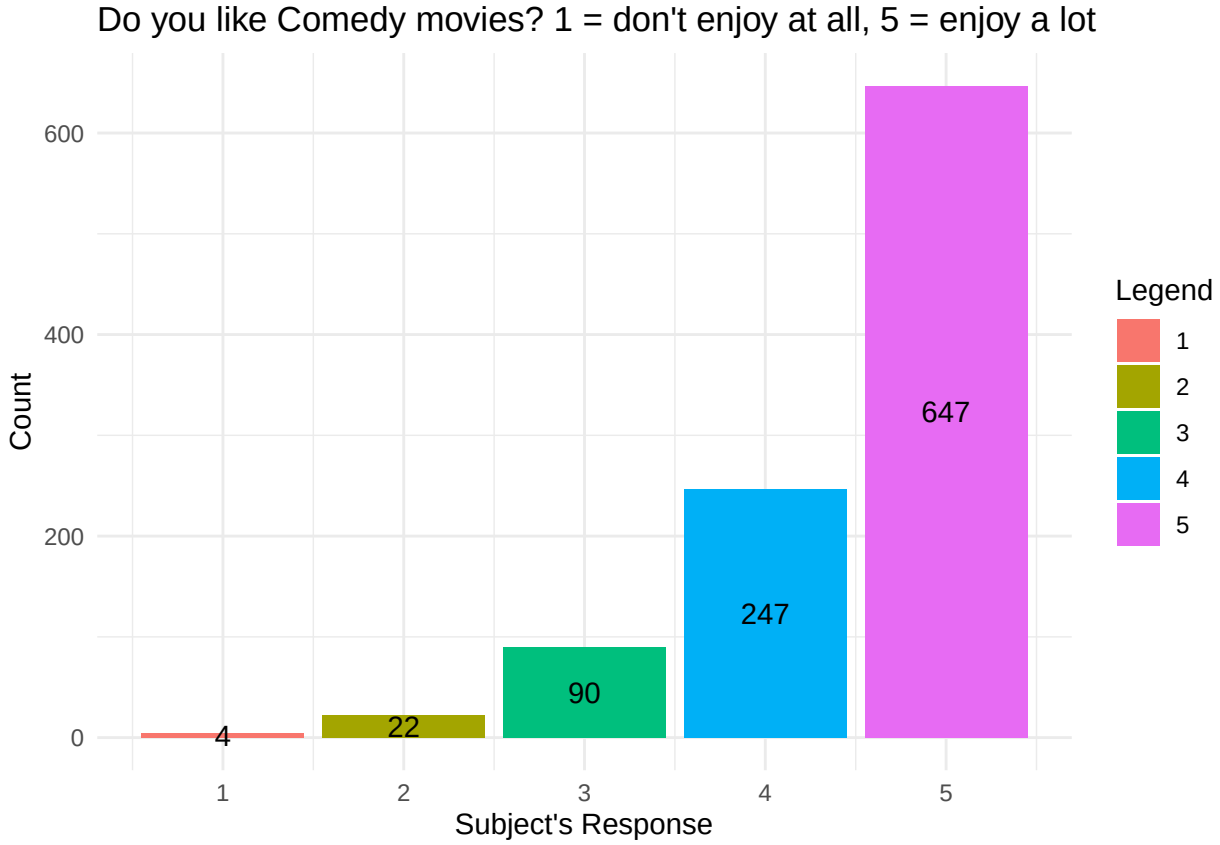
Listed above are the questions for the Movie Preferences section of the survey.

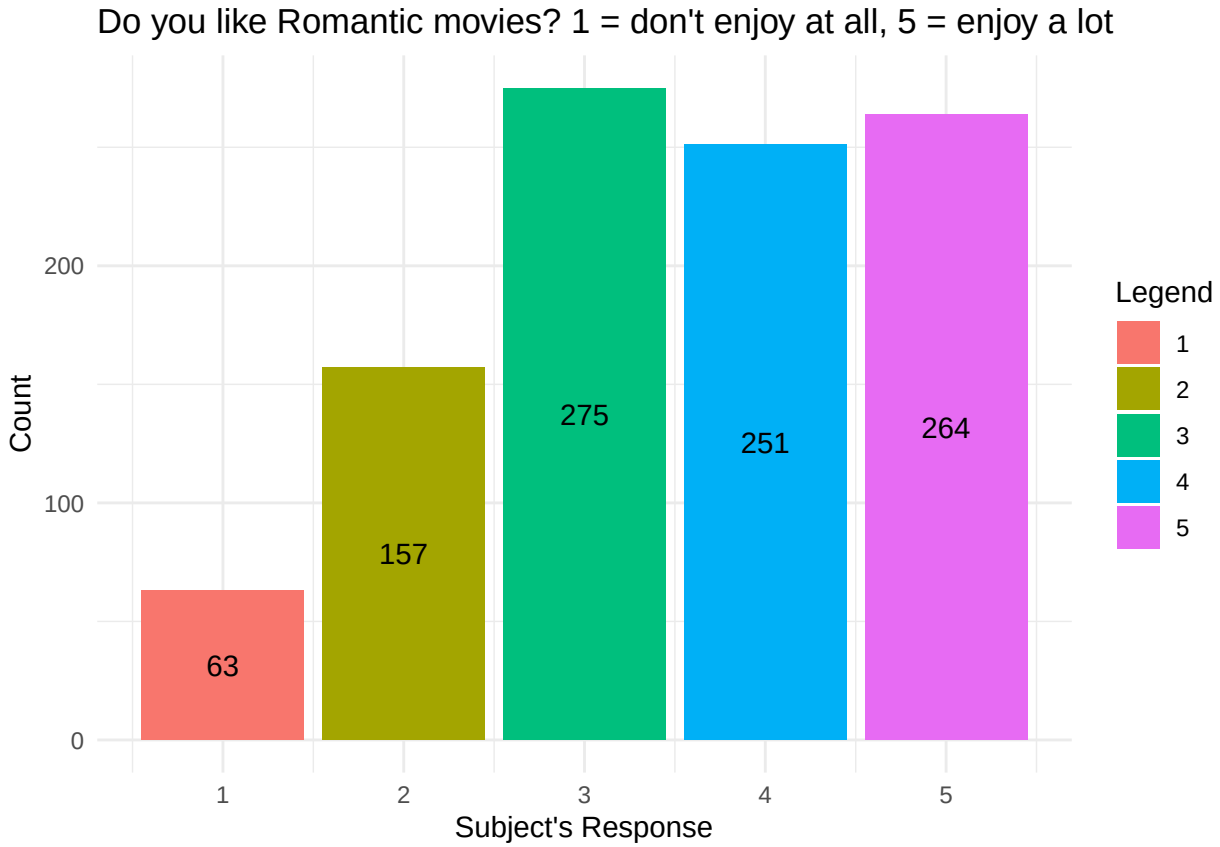
```
# Create vector of movie genres
movie.genres <- colnames(movies.pref[,2:ncol(movies.pref)])

# For loop to create bargraphs showing responses to each movie genre
for(i in 2:ncol(movies.pref)){

plots <- ggplot(data.frame(movies.pref), aes(x = movies.pref[,i],
                                             fill = factor(movies.pref[,i]))) +
  geom_bar() + theme_minimal() +
  geom_text(stat = "count", aes(label = ..count.., y = ..count..),
            position = position_stack(0.5)) +
```

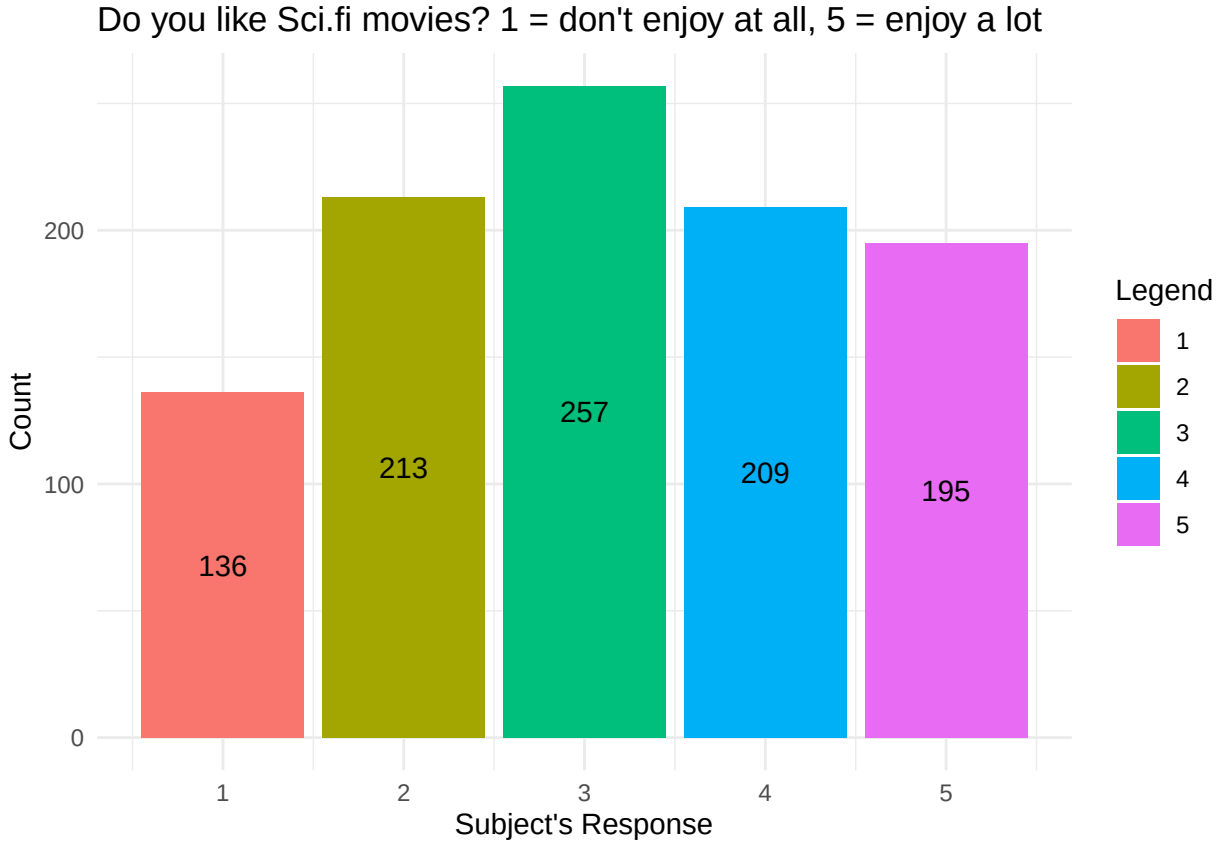
```
  labs(title = paste("Do you like", movie.genres[i-1],
                     "movies? 1 = don't enjoy at all, 5 = enjoy a lot"),
       x = "Subject's Response",  y = "Count", fill = "Legend")

par(mfrow = c(6,3))
print(plots)
}
```

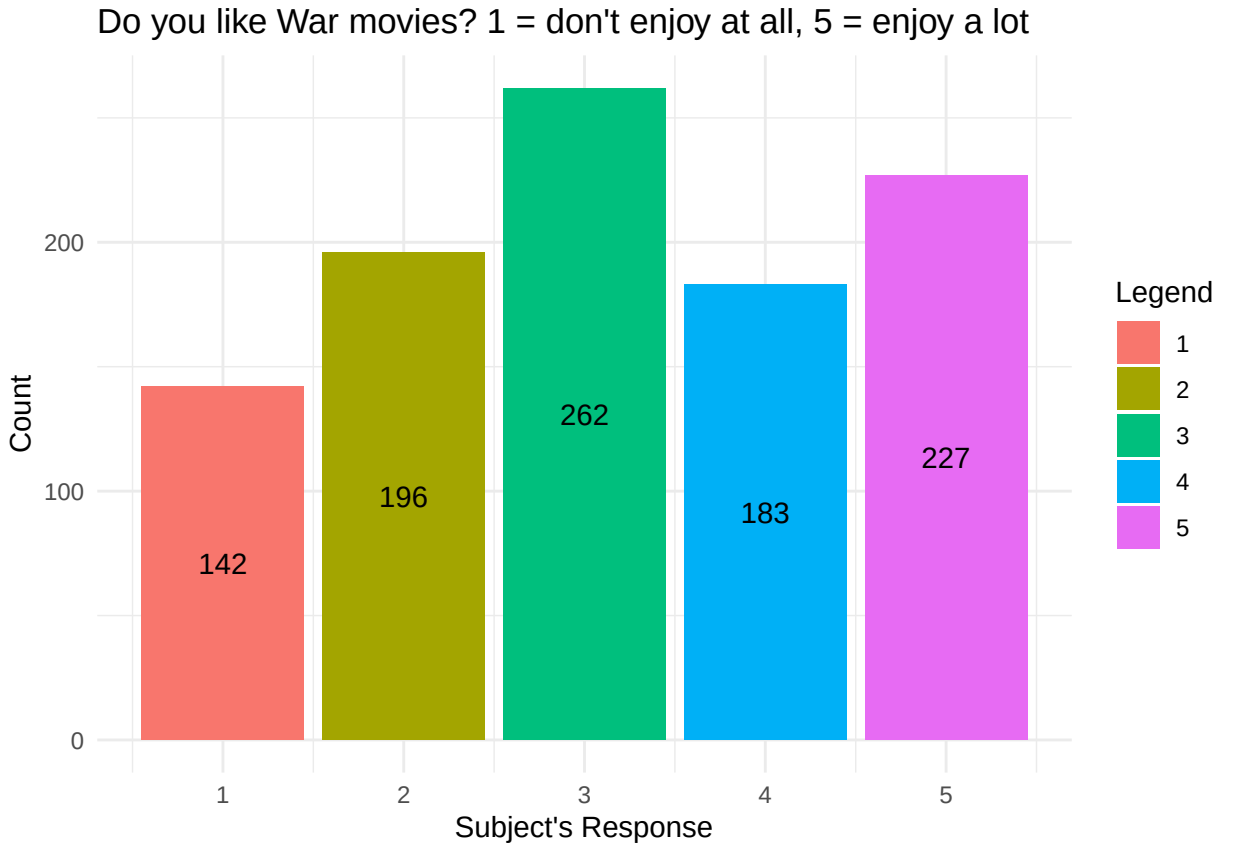


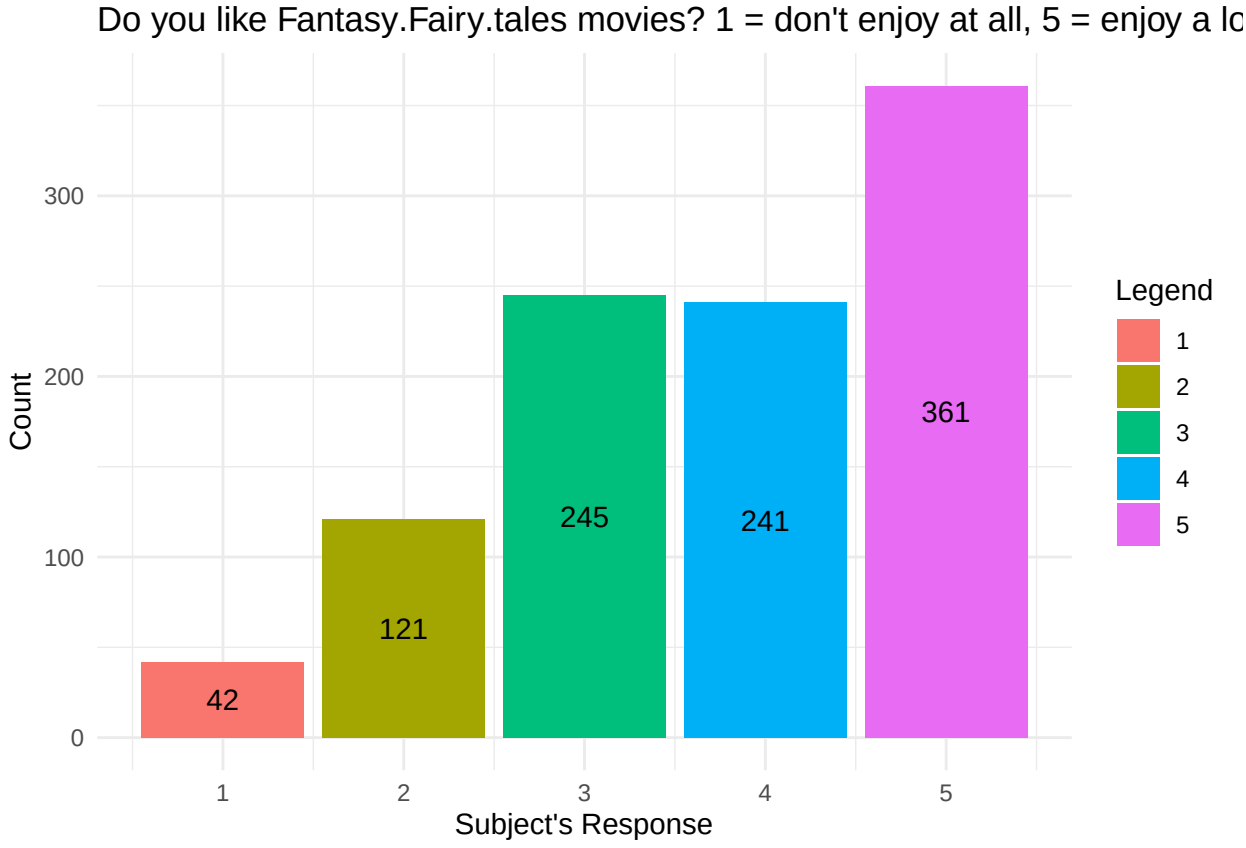

Do you like Thriller movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
77
173
245
315
200
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

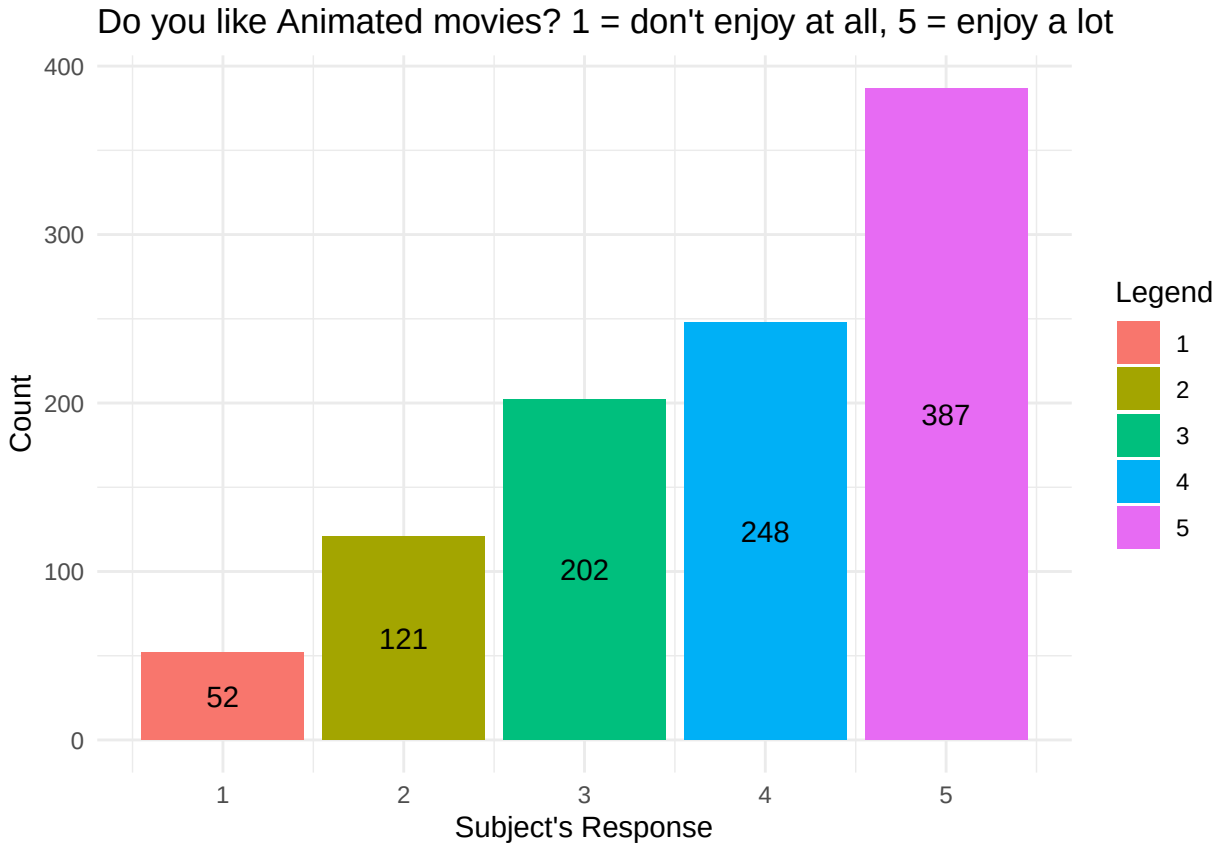

Do you like Comedy movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
600
400
200
0
4
22
90
247
647
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

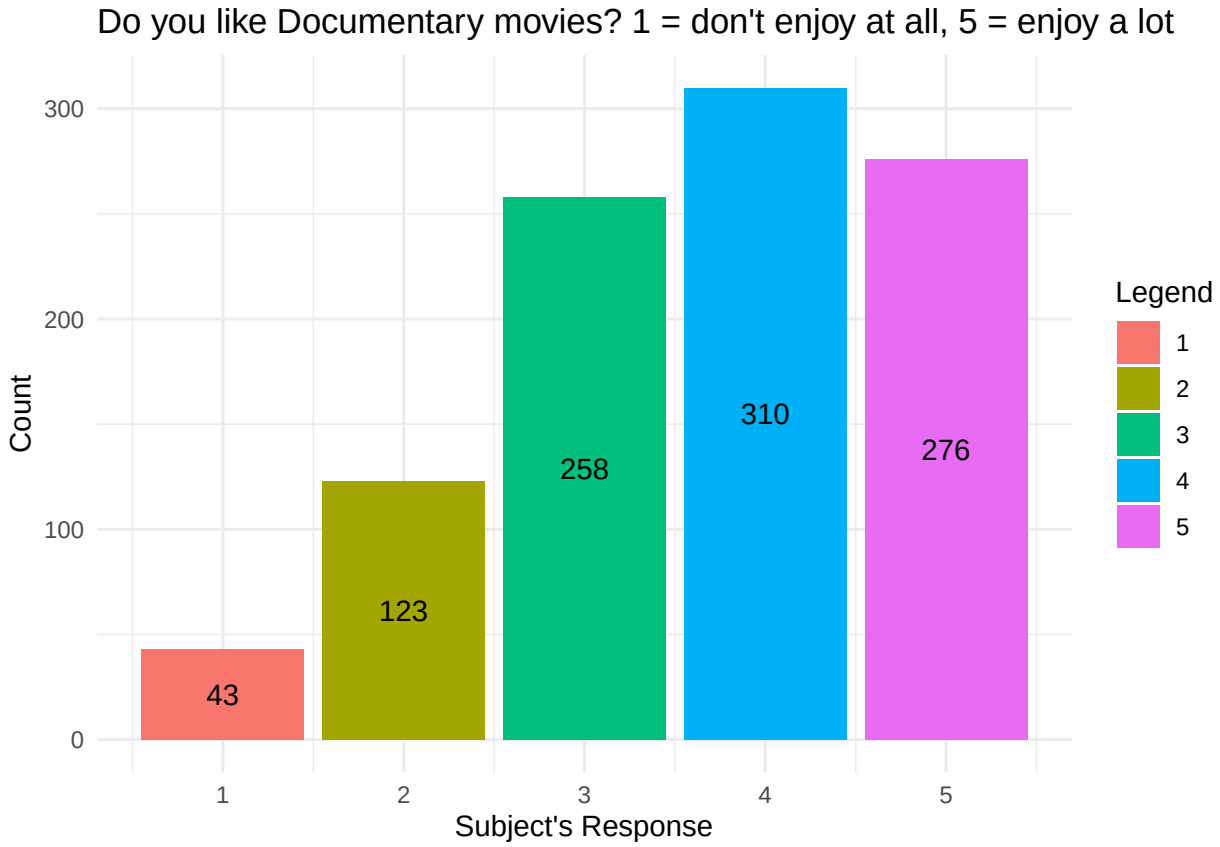

Do you like Romantic movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
200
100
0
63
157
275
251
264
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

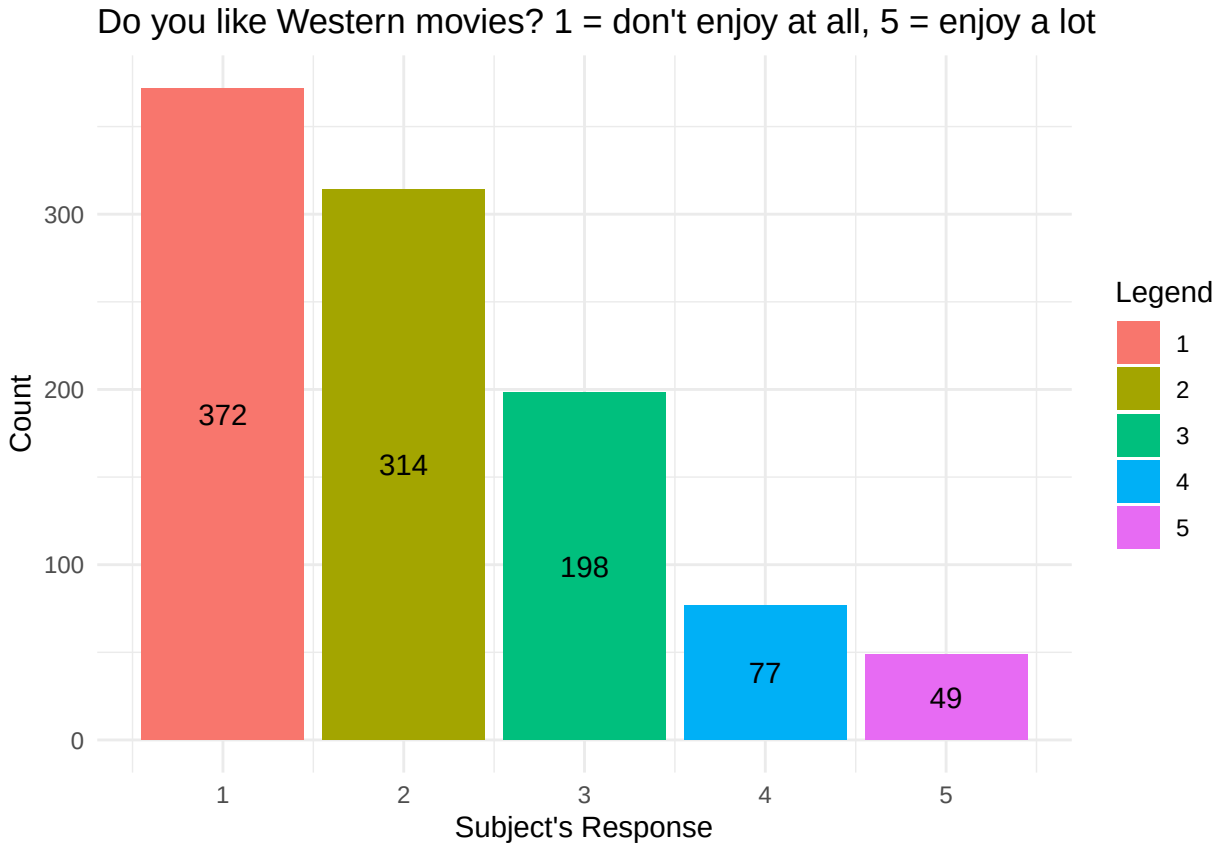

Do you like Sci.fi movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
200
100
0
136
213
257
209
195
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

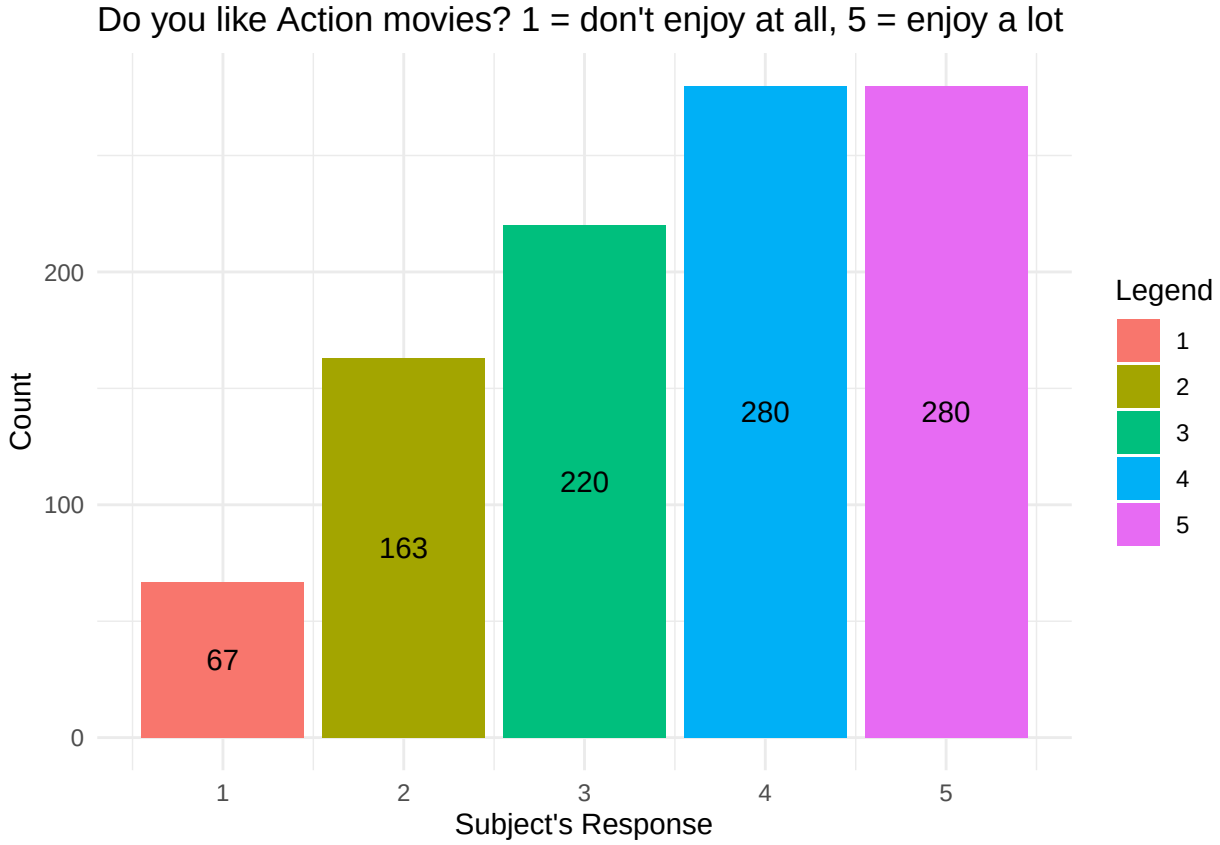

Do you like War movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
200
100
0
142
196
262
183
227
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Fantasy.Fairy.tales movies? 1 = don't enjoy at all, 5 = enjoy a lo
Count
300
200
100
0
42
121
245
241
361
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Animated movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
400
300
200
100
0
52
121
202
248
387
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Documentary movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
43
123
258
310
276
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Western movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
300
200
100
0
372
314
198
77
49
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5


Do you like Action movies? 1 = don't enjoy at all, 5 = enjoy a lot
Count
200
100
0
67
163
220
280
280
1
2
3
4
5
Subject's Response
Legend
1
2
3
4
5

Looking at the bar plots above, we notice:

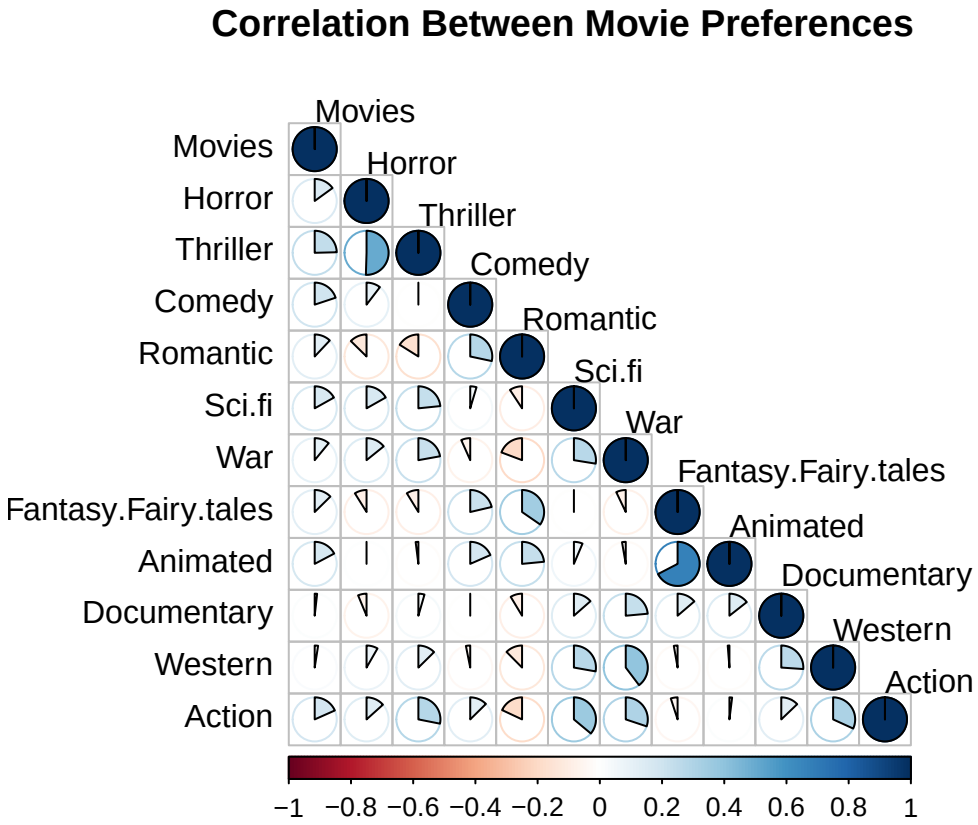
Comedy, Fantasy, and Animated movies are the most popular types of movies

Horror and Western movies are the least liked types of movies

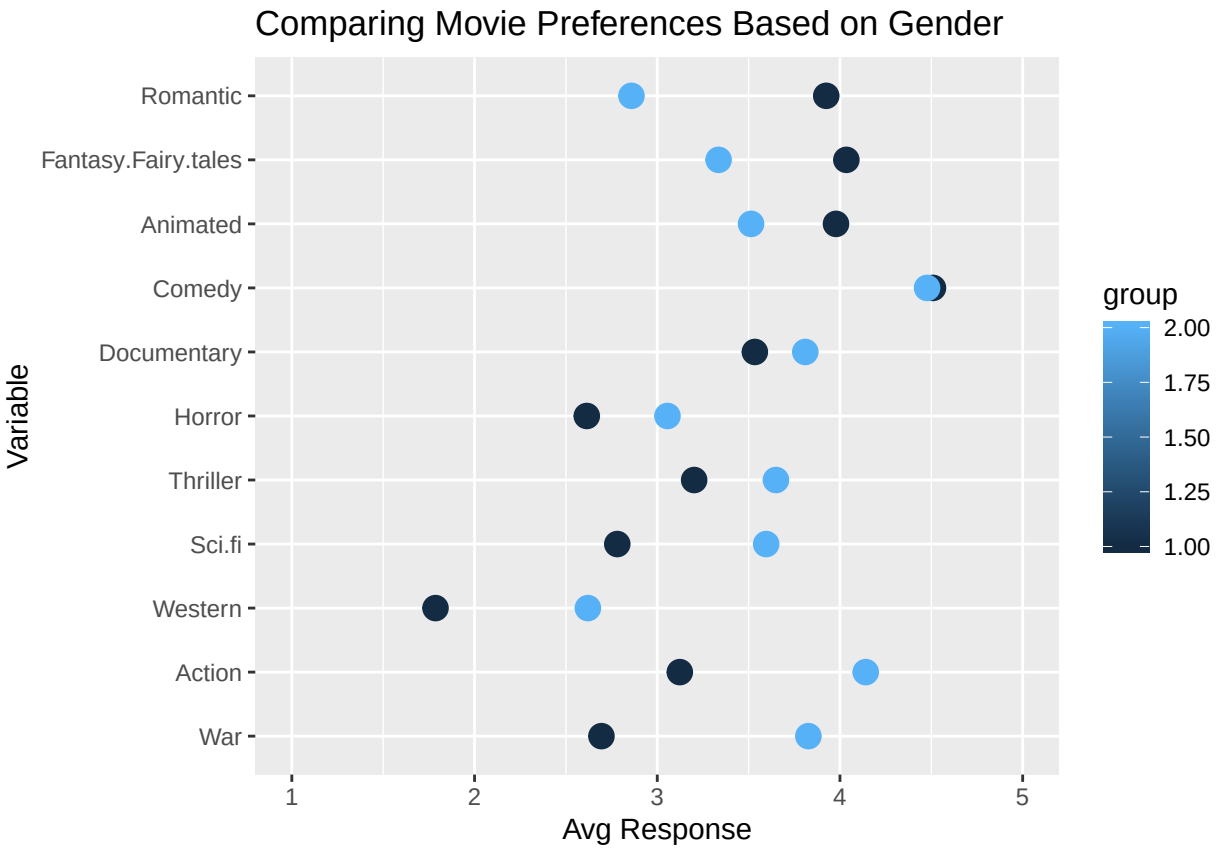
Comedy movies seem to be the most popular genre, with well over half of the particpants (647) saying they enjoy them a lot.

Let's now take a look at the correlation plot of the movie preferences as well as the gender-based preferences.

```
# Correlation plot of movie preferences
corrplot(cor(movies.pref), type = "lower", method = "pi", tl.col = "black",
         tl.srt = .20, title = "Correlation Between Movie Preferences",
         mar=c(0,0,1,0))
```




```
# Run gender difference function on movie section of data
filled.data %>% gender.diff("Gender", "Horror", "Action") +
  ggtitle("Comparing Movie Preferences Based on Gender")
```



Looking at the first of these two graphs, the correlation matrix, we see:

Those who like Animated movies also commonly tend to like Fantasy/Fairy Tale movies.

Those who like Thrillers also tend to like Horror movies

It seems that most of the participants' movie preferences partially depend on whether or not they enjoy Romantic movies because it is that column that has most of the relevant negative correlations. For example, if the participant likes Romantic movies, they tend not to like Sci-Fi, War, Western, or Action movies.

Looking at the second visualiztion:

We can see that movie prefernces vary wildly based on gender.

Males seem to enjoy watching Action and War movies much more than females, while females tend to enjoy Romantic and Fantasy movies much more than males.

Also, Western movies seems to be by far the least favorite type of movie amongst females, while Romantic movies are the least favorite of males.

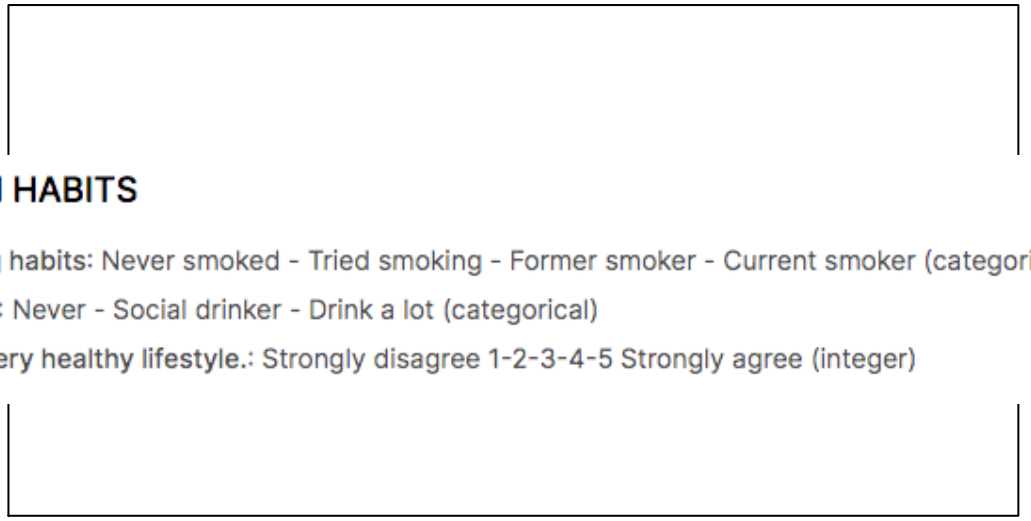
We can also see that people enjoy Comedy movies regardless of gender, confirming what the bar graphs from above showed us.

Health Habits

The next topic on the survey is called Health Habits. This section of the survey contained three questions, which can be viewed below.

```
# Load screenshot of questions on health habits
library(png)
img <- readPNG("HealthHabits.png")
```

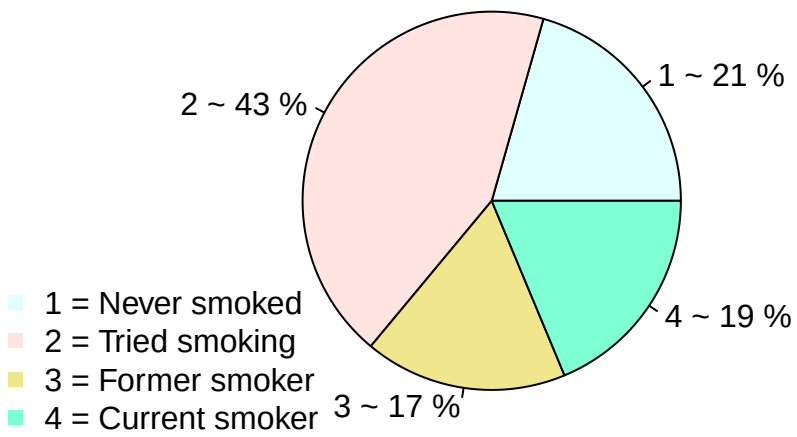
```
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n",
     yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img)
```



Note: While the survey doesn't specify what kind of smoking they're asking about, I will work under the assumption that it's asking about smoking tobacco (cigarettes).

I'll start examining these questions by creating graphics to see the distrubution of responses to each question.

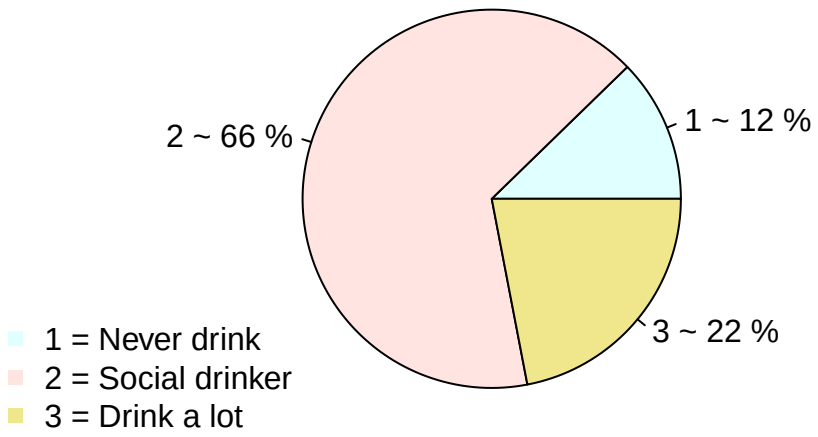
```
# Create pie chart for smoking habits
slices <- as.vector(count(health[,1])$freq)
lbls <- levels(as.factor(health[,1]))
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls,pct, sep = " ~ ")
lbls <- paste(lbls,"%", sep = " ")
colors <- c("lightcyan", "mistyrose", "khaki", "aquamarine")
pie(slices, labels = lbls, col = colors,
    main = "Pie Chart of Smoking Habits (Question 1)")
legend("bottomleft", legend  = c("1 = Never smoked","2 = Tried smoking",
                                 "3 = Former smoker", "4 = Current smoker"),
       col = c("lightcyan", "mistyrose", "khaki", "aquamarine"),
       pch = 15, bty = "n")
```



Looking at the pie chart above, we can see that the most common response to the question of smoking was "Tried Smoking", shown as a response of '2', with 43% of participants marking this as their answer. 21% of the participants said they've never smoked, 19% said they are current smokers, then the other 17% said they were former smokers. Combining these numbers, we see that a large majority, 79%, of the participants have smoked at least once in their lifetime.

Let's see if alcohol is just as prevelant in the lives of this Slovakian population by examining the responses to the second question of this health habits section.

```
# Create pie chart for drinking habits
slices <- as.vector(count(health[,2])$freq)
lbls <- levels(as.factor(health[,2]))
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls,pct, sep = " ~ ")
lbls <- paste(lbls,"%", sep = " ")
colors <- c("lightcyan", "mistyrose", "khaki")
pie(slices, labels = lbls, col = colors,
    main = "Pie Chart of Drinking Habits (Question 2)")
legend("bottomleft",
       legend  = c("1 = Never drink","2 = Social drinker", "3 = Drink a lot"),
       col = c("lightcyan", "mistyrose", "khaki"), pch = 15, bty = "n")
```

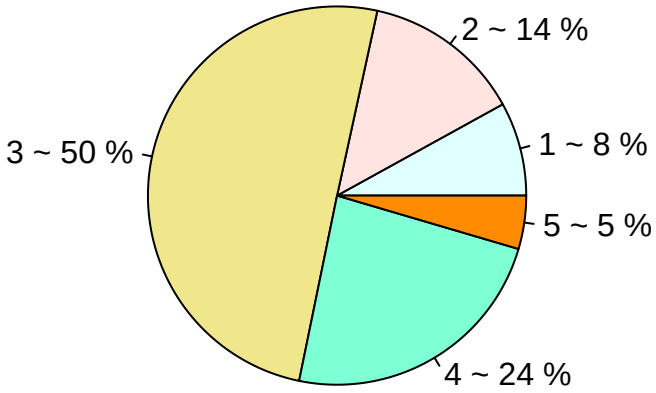


Looking at this pie chart, we see that the most common answer to the question about drinking habits is "Social drinker" with 66% of the people choosing that response. 22% drink a lot and the remaining 12% have never drank. So it can be concluded that 88% of this subpopulation consume alcohol.

Comparing the drinking and smoking habits, it seems as though alcohol is the more common drug of choice amongst this subpopulation in Slovakia, although well over half smoke as well.

Let's see how these findings relate to these people's perception of their overall health.

```
# Create pie chart for lifestyle health
slices <- as.vector(count(health[,3])$freq)
lbls <- levels(as.factor(health[,3]))
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls,pct, sep = " ~ ")
lbls <- paste(lbls,"%", sep = " ")
colors <- c("lightcyan", "mistyrose", "khaki", "aquamarine", "darkorange")
pie(slices, labels = lbls, col = colors,
    main = "Do the Particpants Think They Live a Healthy Life? (Question 3)")
```

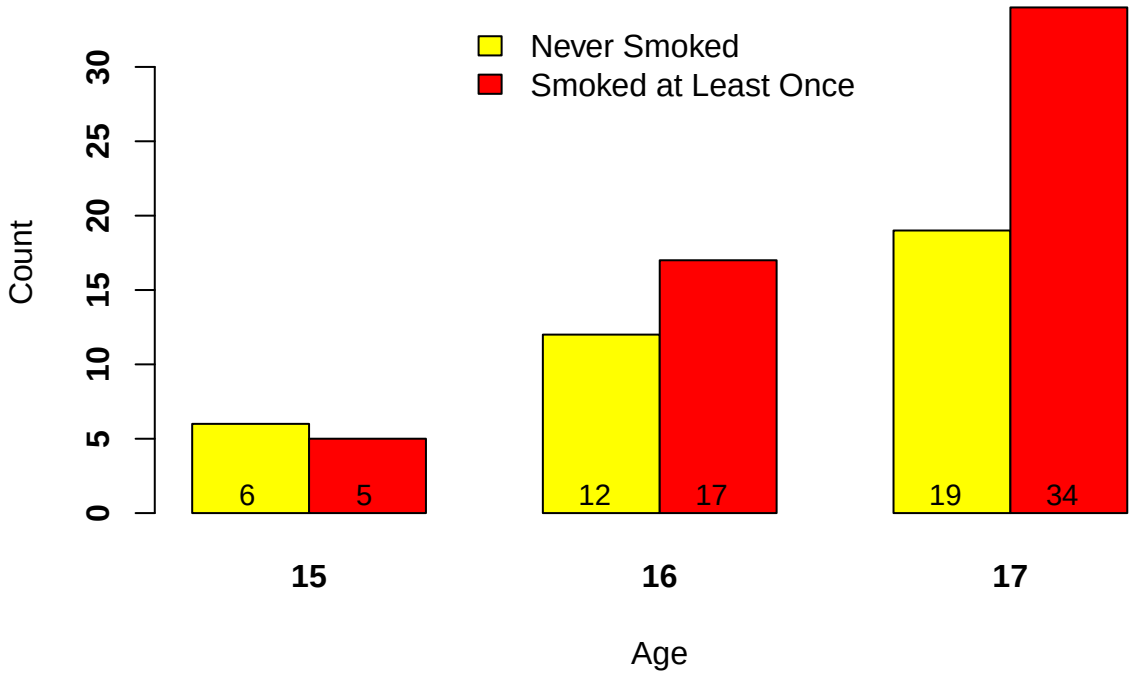


From this pie chart, we see that the most popular response by a large margin was a 3 on the scale from one to five on how much they agree with the statement, "I live a very healthy lifestyle". A '3' response can be interpreted as indifferent or "neither agree nor disagree" with the statement in question. This could be because people don't want to admit to not living a healthy lifestyle. This result makes sense after seeing how prevelant smoking and drinking are in these peoples' lives. It's likely that they don't want to think that using these drugs means they're not living a healthy lifestyle, or they just don't comfortable admitting as much on a survey.

Another interesting thing we can look at using this section of data is whether or not underage use of drugs is an issue in Slovakia. A quick google search tells us that the legal drinking age in Slovakia is 21 while the legal smoking age is 18. Each participant entered their age on the survey as observed earlier in this report, so let's take a look at the use of these drugs within the age groups below the respective legal limits.

I'll start with smoking. Since a response of '1' on the survey is the only one saying the participant has never smoked, I'll look at those responses compared to the other three combined (tried smoking, former smoker, and current smoker). This will allow us to look at those who have smoked at least once in their lifetime compared to those who have not.

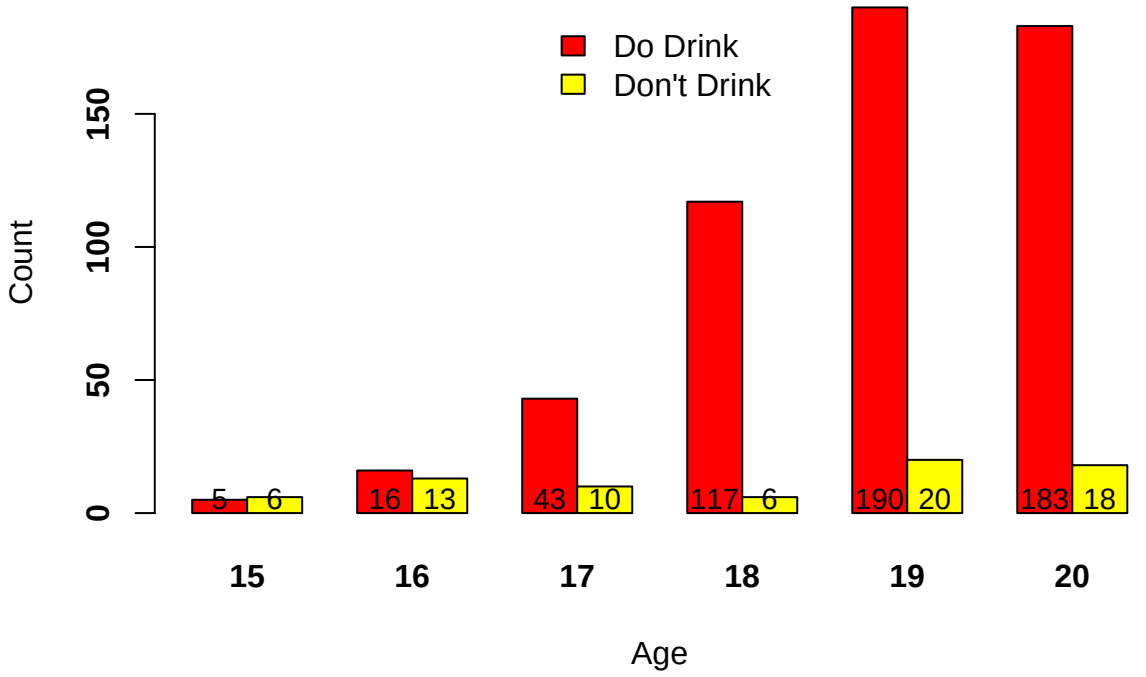
```
# Recode smoking variable so it's just haven't smoked vs. have smoked
x <- filled.data$Smoking
x <- recode(x, '1' = "Never Smoked", '2' = "Smoked at Least Once",
            '3' = "Smoked at Least Once", '4' = "Smoked at Least Once")

# Create table showing smoking habits per age group, then plot it
smoking.tbl <- table(x, filled.data$Age)[,1:3]
smoking.bp <- barplot(smoking.tbl, beside = T, legend.text = T,
                      col = c("yellow", "red"),
                      xlab = "Age", ylab = "Count", font.axis = 2,
                      main = "Observing if Underage Smoking is an Issue in Slovakia",
                      args.legend = list(x = "top", bty = "n",
                                         inset=c(-0.15, 0)))
text(smoking.bp, 0, smoking.tbl, adj = c(.7,-.4), cex = .9, col = "black")
```



As stated before, the legal smoking age in Slovakia is 18 so I limited the bargraph to the ages 15 through 17. As seen in the bargraph above, the majority of the underage participants in this survey have smoked at least once in their life. This shows that, just like in many other parts of the world including America, the underage use of tobacco is an issue in Slovakia.

One can assume there is a similar issue with the consumption of alcohol, but let's take a look at the responses to that question in a similar fashion to confirm.

```
# Re-categorize drinking responses so it's just do or don't drink
x2 <- filled.data$Alcohol
x2 <- recode(x2, '1' = "Don't Drink", '2' = "Do Drink", '3' = "Do Drink")

# Create table showing drinking habits per age group and then plot it
drinking.tbl <- table(x2, filled.data$Age)[,1:6]
drinking.bp <- barplot(drinking.tbl, beside = T, legend.text = T, col = c("red", "yellow"),
                       xlab = "Age", ylab = "Count", font.axis = 2,
                       main = "Observing if Underage Drinking is an Issue in Slovakia",
                       args.legend = list(x = "top", bty = "n", inset=c(-0.15, 0)))
text(drinking.bp, 0, drinking.tbl, adj = c(.5,-.2), cex = .9, col = "black")
```



Since the legal drinking age in Slovakia is older than the legal smoking age, this bargraph observes the age groups from 15 to 20. As seen in this bargraph, the number of participants who drink far outnumbers those who don't drink for every age group except for 15. We can see that underage drinking is a much bigger problem in Slovakia than smoking is. This is most likely due to the fact that they can legally start smoking at 18 whereas the law doesn't allow them to drink until 21, so the phrase 'underage consumption' can be applied for a few more years. But the fact still stands that underage drinking in Slovakia is clearly very common, much like the rest of the world.

Personality Traits

The next section of the survey asked questions about various personality traits. I'm interested to see if there is any relation between one trait and the next, or in other words, if having one personality trait makes one more likely or less likely to have another. I will do this using a correlation matrix.

But first, here is a list of the questions for this section of the survey to gain a better understanding of any results that come after.

```
# Load screenshot of personality traits questions from survey
img <- readPNG("Pers1.png")
img2 <- readPNG("Pers2.png")
img3 <- readPNG("Pers3.png")
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n", yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img)
```

PERSONALITY TRAITS, VIEWS ON LIFE & OPINIONS

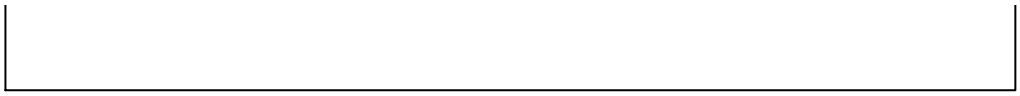
1. I take notice of what goes on around me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. I try to do tasks as soon as possible and not leave them until last minute.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. I always make a list so I don't forget anything.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. I often study or work even in my spare time.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. I look at things from all different angles before I go ahead.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. I believe that bad people will suffer one day and good people will be rewarded.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. I am reliable at work and always complete all tasks given to me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
8. I always keep my promises.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
9. I can fall for someone very quickly and then completely lose interest.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
10. I would rather have lots of friends than lots of money.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
11. I always try to be the funniest one.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
12. I can be two faced sometimes.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
13. I damaged things in the past when angry.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
14. I take my time to make decisions.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
15. I always try to vote in elections.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
16. I often think about and regret the decisions I make.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
17. I can tell if people listen to me or not when I talk to them.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
18. I am a hypochondriac.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
19. I am emphatetic person.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
20. I eat because I have to. I don't enjoy food and eat as fast as I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
21. I try to give as much as I can to other people at Christmas.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
22. I don't like seeing animals suffering.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
23. I look after things I have borrowed from others.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
24. I feel lonely in life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

```
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n", yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img2)
```

24. I feel lonely in life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
25. I used to cheat at school.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
26. I worry about my health.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
27. I wish I could change the past because of the things I have done.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
28. I believe in God.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
29. I always have good dreams.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
30. I always give to charity.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
31. I have lots of friends.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
32. Timekeeping.: I am often early. - I am always on time. - I am often running late. (categorical)
33. Do you lie to others?: Never. - Only to avoid hurting someone. - Sometimes. - Everytime it suits me. (categorical)
34. I am very patient.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
35. I can quickly adapt to a new environment.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
36. My moods change quickly.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
37. I am well mannered and I look after my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
38. I enjoy meeting new people.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
39. I always let other people know about my achievements.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
40. I think carefully before answering any important letters.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
41. I enjoy childrens' company.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
42. I am not afraid to give my opinion if I feel strongly about something.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
43. I can get angry very easily.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
44. I always make sure I connect with the right people.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
45. I have to be well prepared before public speaking.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
46. I will find a fault in myself if people don't like me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
47. I cry when I feel down or things don't go the right way.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
48. I am 100% happy with my life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
49. I am always full of life and energy.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

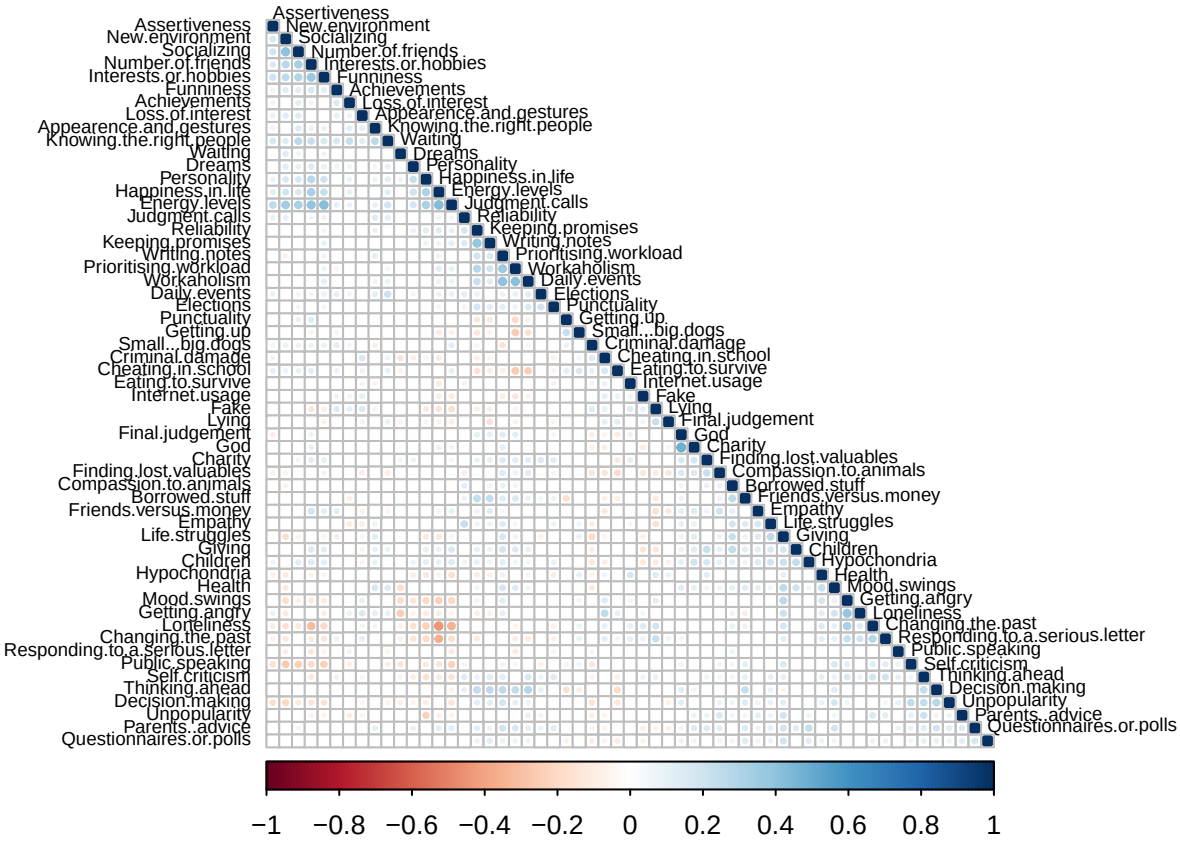
```
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n", yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img3)
```

50. I prefer big dangerous dogs to smaller, calmer dogs.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
51. I believe all my personality traits are positive.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
52. If I find something the doesn't belong to me I will hand it in.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
53. I find it very difficult to get up in the morning.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
54. I have many different hobbies and interests.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
55. I always listen to my parents' advice.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
56. I enjoy taking part in surveys.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
57. How much time do you spend online?: No time at all - Less than an hour a day - Few hours a day - Most of the day (categorical)



```
# Get p-values for each correlation coefficient
cor.mtest <- function(mat, ...) {
    mat <- as.matrix(mat)
    n <- ncol(mat)
    p.mat<- matrix(NA, n, n)
    diag(p.mat) <- 0
    for (i in 1:(n - 1)) {
```

```
        for (j in (i + 1):n) {
            tmp <- cor.test(mat[, i], mat[, j], ...)
            p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
        }
    }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}
# Create correlation matrix, excluding those with insignificant p-values
p.mat <- cor.mtest(personality)
corrplot(cor(personality), type = "lower", method = "circle", order = "hclust",
         tl.col = "black", tl.cex = .6, tl.srt = .45, p.mat = p.mat,
         sig.level = .05, insig = "blank")
```



This is a correlation graphic displaying the relations between personality traits, but while also taking into account the degree or 'level' of relation. A p-value was calcualted for each correaltion coefficient to determine if it could be deemed statistically significant at a signifiance level of .05. If it was deemed insignificant, that box was left empty. So all the boxes with circles remaining in them mean the relation between that set of personality traits is significant (with blue circles showing a positive correlation and orange circles showing negative correlation). Looking at this, we notice:

A person's energy levels is positively correlated with many other aspects of their personality. For example, higher energy levels leads to more interests and hobbies, more friends, and the ability to better adapt to new environments. It even leads to an overall higher feeling of happiness in life.

Another interesting and quite notable positive correlation seen above is the relation between God and Charity. This visual shows us that a belief in God generally leads to more donations to charities among this group of Slovakians.

As for negative correlations, the most prevalent one is Loneliness which has a negative effect on several other personality traits. If a person said they felt lonely, they generally also noted a lack of energy, held a belief that their personality was not becoming to others, and also felt a general lack of happiness in their life.

These are just a few of the more notable insights we can observe from this one visualization of the responses to the personality section of this survey. It's very interesting and important to know how all facets of a person's personality effect the rest of their persona, which then allows humans to become more empathetic and try to help those in need.

Spending Habits

The final section of this survey covered spending habits. My interest in this section is to observe how one's age effects their decision making as far as what they spend their money on.

Here is the list of questions for this section:

```
# Load screenshot of spending habit questions from survey
img <- readPNG("Spending.png")
plot(NA,xlim=c(0,2), ylim = c(0,4), type = "n", xaxt = "n", yaxt = "n", xlab = "", ylab = "")
grid::grid.raster(img)
```

I will study the spending habits based on age by creating a scatterplot. For each age along the x-axis, there will be several points on the y-axis which will show the mean response of that age group to each of the questions above.

```
# Create data frame with ages and spending habit responses
spending.by.age <- cbind(filled.data$Age, spending)

# Rename age column to easily extract later
spending.by.age <- rename(spending.by.age, c("filled.data$Age" = "Age"))

# Calculate mean responses to each question for each age group
spend.info <- aggregate(spending.by.age[,2:ncol(spending.by.age)], list(spending.by.age$Age), mean)
plot(spend.info$Group.1,spend.info$Finances, ylim = c(2.2,4.75),
     col = "burlywood", pch = 19, cex = 1.1,
     xlab = "Age", ylab = "Mean Response")
points(spend.info$Group.1,spend.info$Shopping.centres, col = "firebrick2", pch = 19, cex = 1.1)
points(spend.info$Group.1, spend.info$Branded.clothing, col = "lawngreen", pch = 19, cex = 1.1)
points(spend.info$Group.1, spend.info$Entertainment.spending, col = "gold", pch = 19, cex = 1.1)
```

```
points(spend.info$Group.1, spend.info$Spending.on.looks, col = "purple", pch = 19, cex = 1.1)
points(spend.info$Group.1, spend.info$Spending.on.gadgets, col = "black", pch = 19, cex = 1.1)
points(spend.info$Group.1, spend.info$Spending.on.healthy.eating, col = "coral4", pch = 19,
       cex = 1.1)
legend(20,4.9, legend = c("Savings","Shopping Centres", "Branded Clothing",
                          "Partying/Socializing","Appearance/Looks",
                          "Gadgets", "Food/Healthy Eating"),
       col = c("burlywood","firebrick2","lawngreen","gold",
               "purple","black","coral4"), cex = .8, pch = 15, bty = "n")
ticks = c(15:30)
axis(side = 1, at = ticks)
```

This scatterplot shows us plenty of things about the spending habits according to age amongst those surveyed:

As far as overall trends:

No matter how old a person is, they generally prioritize spending their money on healthy eating over all else.

Spending on gadgets is uncommon no matter the age.

Spending money at shopping centres is very common as a teenager, then becomes less important to early-twenty year olds, but then more common again amongst the mid-to-late twenties age group.

A few interesting notes of each individual age group:

15 year olds spend the majority of their money at shopping centres with the least of it going towards gadgets.

16 year olds also spend most of their money at shopping centres and don't worry much about saving money (shown by the beige circle).

17 year olds prioritize spending money on healthy food and save little money.

18 through 21 year olds focus on spending money on healthy food and put little money into buying gadgets.

For 22 and 23 year olds, shopping centres are the last places they spend their money.

For 25 year olds, spending money on appearance and branded clothing are near the bottom of their priorities, but for 26 year olds, both those spending habits are near the top of their lists.

Conclusion

This report just demonstrates a few of many ways survey data can be gathered and analyzed to better understand its subjects. Using statistical thoughts and processes to observe human life can be tricky since the former is very black & white while the ladder is full of intricacies. However, key insights can still be obtained using such processes as observed in this report.