

STA404/504 HW4

Ethan Gutknecht

08/21/2022

Learning Objectives:

- Basic visualizations and aesthetic settings

The items below will be considered for grading:

- The plots are correct, with professionalism.
- Axis labels and titles are correct and complete.
- Units are clearly labeled.
- Proper grammar in the write-up.
- The discussion and the story told is interesting and appropriate.

Question 1

Load the shuffledplaylist.csv dataset into R by the code below. This data includes information on songs from a shuffled Spotify playlist.

```
#read the dataset to R-studio
music<-read.csv("shuffled_playlist.csv")
```

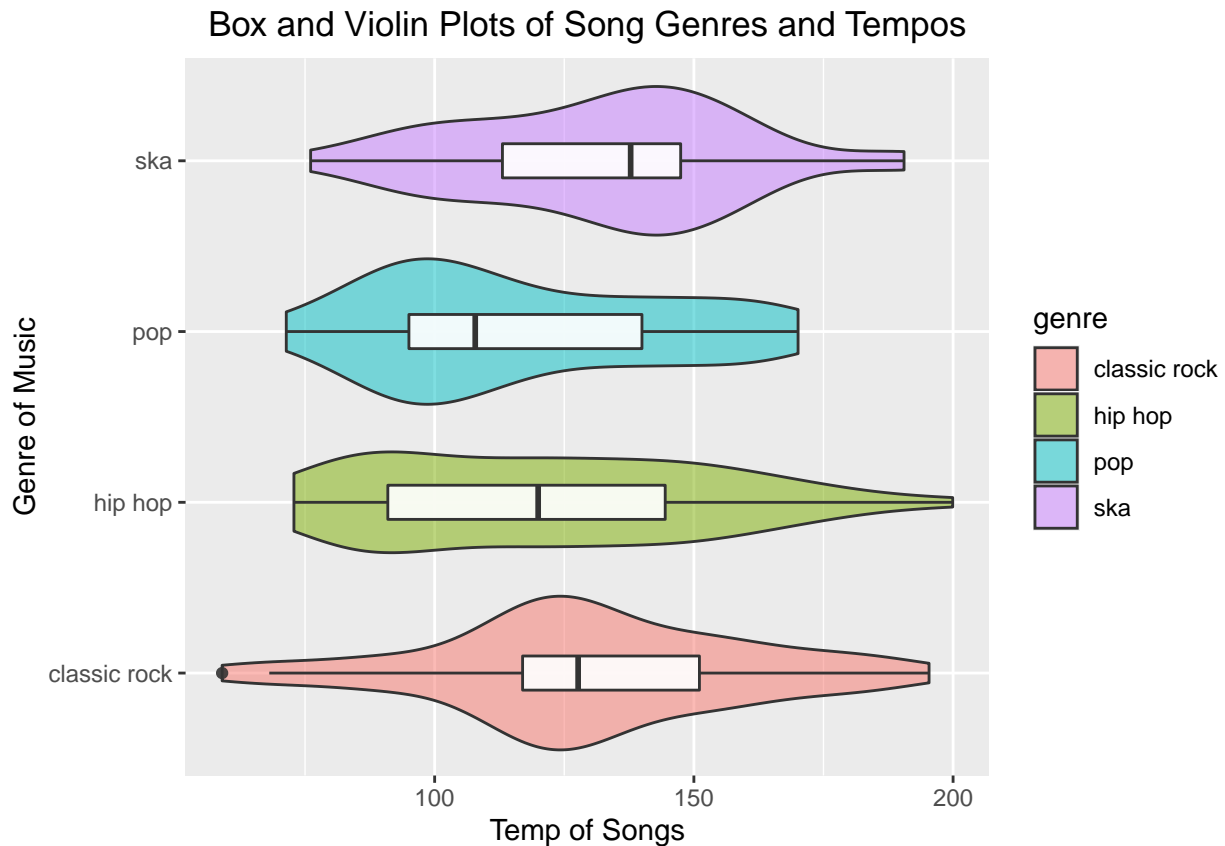
```
# Read data set
music <- read.csv("shuffled_playlist.csv")
head(music)
```

```
##               artist               album               genre
## 1 Streetlight+Manifesto Everything Goes Numb          ska
## 2               Adele                25              pop
## 3      Pink Floyd The Dark Side of the Moon classic rock
## 4 Childish+Gambino  "Awaken, My Love!"          hip hop
## 5      Ed+Sheeran              ÷ (Deluxe)          pop
## 6 Kendrick+Lamar    To Pimp A Butterfly          hip hop
##               name danceability energy loudness speechiness
## 1      A Moment of Violence      0.609 0.9180    -5.998    0.0894
## 2      Water Under the Bridge      0.596 0.8380    -6.520    0.0704
## 3              Speak to Me      0.592 0.0196   -33.350    0.0358
## 4              Baby Boy      0.529 0.4030    -7.103    0.0349
## 5      Nancy Mulligan      0.680 0.8520    -4.350    0.0349
## 6 You Ain't Gotta Lie (Momma Said) 0.501 0.6460    -5.422    0.2310
## acousticness instrumentalness liveness valence  tempo
## 1      0.0087      0.00e+00  0.2940  0.5410  89.975
```

```
## 2      0.0189      1.54e-05  0.1080  0.4790  94.982
## 3      0.3620      8.54e-01  0.1080  0.0322 119.504
## 4      0.1230      4.96e-01  0.0647  0.7230 199.928
## 5      0.1170      0.00e+00  0.0866  0.8520 101.993
## 6      0.4820      1.43e-05  0.2740  0.6480  94.746
```

How do the tempo values differ by genre? Use a side-by-side style of plot that shows both the distribution, as well as summary statistics to help tell a story of how they differ. Describe in a few sentences what this display tells us about the relationship between the variables.

```
# Create plot
ggplot(aes(x=tempo), data=music) +
  geom_violin(aes(y=genre, fill=genre), alpha=0.5) +
  geom_boxplot(aes(y=genre), alpha=0.9, width=0.2) +
  labs(x="Temp of Songs", y= "Genre of Music") +
  ggtitle("Box and Violin Plots of Song Genres and Tempos") +
  theme(plot.title = element_text(hjust = 0.5))
```



```
# Show summary stats
music %>% group_by(genre) %>% summarise(meanTempo=mean(tempo), medianTemp=median(tempo))
```

```
## # A tibble: 4 x 3
##   genre      meanTempo medianTemp
##   <chr>         <dbl>         <dbl>
## 1 classic rock    130.         128.
```

```
## 2 hip hop      121.      120.
## 3 pop          117.      108.
## 4 ska          132.      138.
```

Question 2

The data “**Arthritis**” is a dataset from the package “**vcd**”. It contains data in Koch & Edwards (1988), which is collected from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis. To access the dataset, you may use the following code:

```
install.packages("vcd")
library(vcd)
#Load Arthritis dataset (data frame)
data(Arthritis)
?Arthritis
```

```
# install.packages("vcd")
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.1.3
```

```
## Loading required package: grid
```

```
#Load Arthritis dataset (data frame)
data(Arthritis)
?Arthritis
```

```
## starting httpd help server ...
```

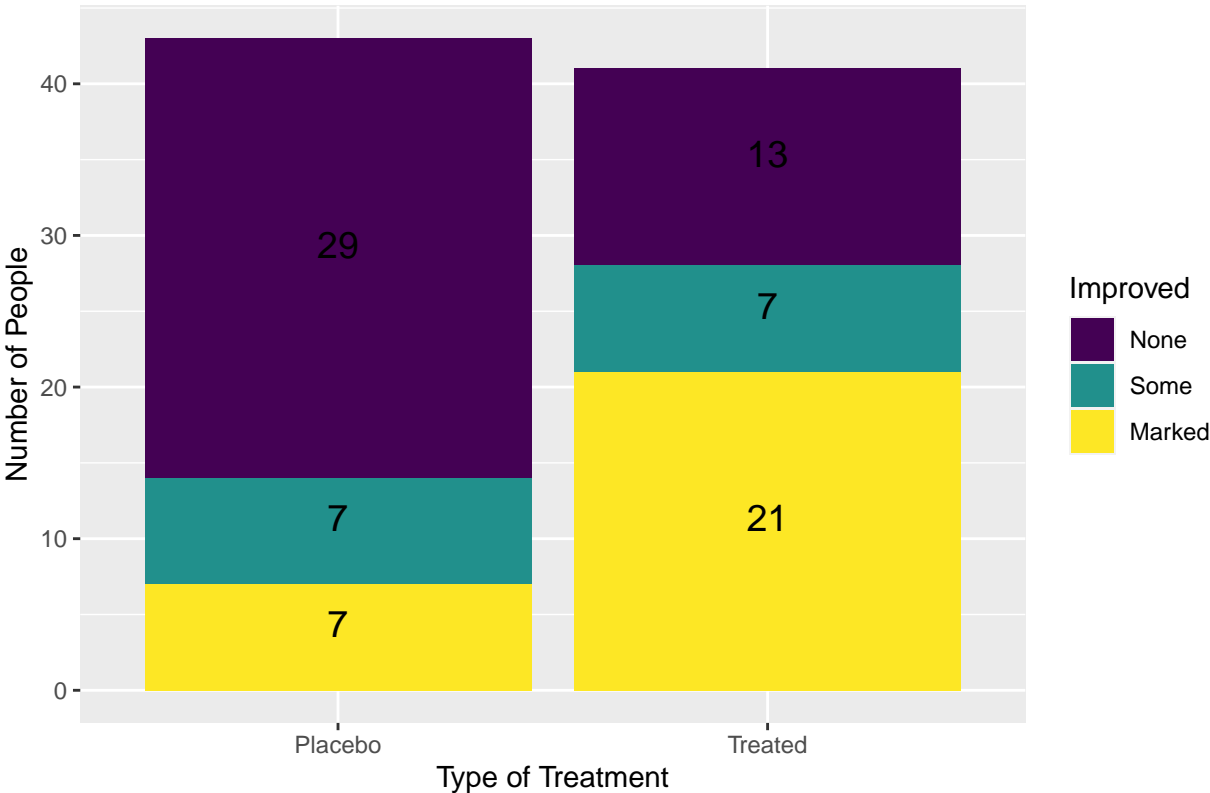
```
## done
```

Please create a stacked/grouped bar graph that summarizes the number of patients who received Placebo/Treatment, and the outcome of the treatment (None, Some, or Marked improvement). On the plot, please have the numbers displayed in the center of the corresponding bar. Make sure appropriate aesthetic settings such as color, labels, titles (make it centered), legends, etc, are used, if applicable.

```
# Stacked bar
# patients who recieved placebo/treatment
# outcome of treatment
# numbers displayed on center
# colors, titles, etc

ggplot() +
  geom_bar(aes(x=Treatment, fill=Improved), data=Arthritis, stat="count",
           position="stack") +
  geom_text(aes(x=Treatment, label=..count.., group=Improved), stat="count",
           data=Arthritis, position = position_stack(vjust= 0.5), size=5, vjust=-0) +
  labs(x="Type of Treatment", y="Number of People") +
  ggtitle("Stacked-Bar Graph of Treatments of Arthritis Patients and the Outomes") +
  theme(plot.title = element_text(hjust = 0.5))
```

Stacked-Bar Graph of Treatments of Arthritis Patients and the Outcomes



Based on my graph, you can see the treated patients had more people marked as improved than the placebo.