Statistical explorations, data preparation, and correlations

Andri Setiyawan Benedikt Meyer Niri Gala
Yosep Dwi Kristanto

```
library(tidyverse)
library(haven)
library(labelled)
library(rvest)
library(knitr)
library(kableExtra)
```

What is your favorite song? How old is the song relative to your year of birth? Do you think others might have similar answers to the latter question?

Through this report, we will present the results of our investigation into the relationship between the age of a song and its average rating. The age of the song we refer to is the difference between the release year of a song and the year of birth of the person listening to that song. The dataset we used is provided by a study published in *Marketing Letters*.

The study surveyed 1,036 residents of the US aged between 18 and 84 years. They were asked to listen to and rate 34 songs. All the songs were selected from the Billboard Top 10 charts from 1950 to 2016. Table 1 presents the list of songs.

```
# Read the webpage
url <- "https://link.springer.com/article/10.1007/s11002-022-09626-7/tables/2"
webpage <- read_html(url)

# Extract the table
music_table <- webpage |>
  html_node("table") |>
  html_table() |>
  rename(
    song_year = 'Song Year',
```

```
    song_title = 'Song Title',
    performers = 'Performer/s'
  )
music_table |>
  kbl() |>
  kable_styling(
    full_width = TRUE,
    bootstrap_options = c("striped", "condensed")
  )
```

Data preparation

The dataset we used initially had 74 variables. We added two variables, `time_music_last_week` and `rating_avg`, using the existing variables. A summary of the 76 variables formed is presented in Table 2.

```
# Import data
hw2_data <- read_sav("11002_2022_9626_MOESM1_ESM.sav") |>
  mutate(
    across(everything(), ~ na_if(., 99))
  ) |>
  mutate(
    time_music_last_week = Q6ax1_1 + Q6ax2_1 / 60,
    Q19_avg = rowMeans(across(starts_with("Q19_1_")), na.rm = TRUE)
  )

# Make a table of variable names and labels
variable_names <- names(hw2_data)
variable_labels <- sapply(hw2_data, var_label)
variable_table <- tibble(
  variable = variable_names,
  label = variable_labels
)
variable_table |>
  kbl() |>
  kable_styling(
    full_width = TRUE,
    bootstrap_options = c("striped", "condensed")
  )
```

Table 1: Music stimuli used in the survey and their respective year in the Billboard charts

song_year	song_title	performers
1950	Play a Simple Melody	Bing and Gary Crosby
1952	You Belong to Me	Jo Stafford
1954	Sh Boom Sh Boom	The Crew Cuts
1956	My Prayer	The Platters
1958	Patricia	Perez Prado
1960	Running Bear	Johnny Preston
1962	Roses are Red	Bobby Vinton
1964	I Get Around	Beach Boys
1966	The Last Train to Clarksville	The Monkees
1968	People Got to be Free	The Rascals
1970	Raindrops Keep Fallin' on My Head	B.J. Thomas
1972	Lean on Me	Bill Withers
1974	The Sound of Philadelphia	MFSB ft. Three Degrees
1976	Play that Funky Music	Wild Cherry
1978	Stayin' Alive	Bee Gees
1980	Crazy Little Thing Called Love	Queen
1982	Don't You Want Me	Human League
1984	Footloose	Kenny Loggins
1986	Party All the Time	Eddie Murphy
1988	Sweet Child O' Mine	Guns N' Roses
1990	Vogue	Madonna
1992	Under the Bridge	Red Hot Chilli Peppers
1994	All She Wants	Ace of Base
1996	Missing	Everything but the Girl
1998	Crush	Jennifer Paige
2000	Say My Name	Destiny's Child
2002	Dilemma	Nelly ft. Kelly Rowland
2004	Hey Ya	OutKast
2006	Sexy Back	Justin Timberlake
2008	Lollipop	Lil Wayne
2010	California Gurls	Katy Perry
2012	Payphone	Maroon 5
2014	Counting Stars	One Republic
2016	Work	Rihanna

Table 2: Summary of variables and their labels

variable	label
S1	S1 - Have you listened to music of your choosing in the last week?
Q1	Q1 - What year were you born in? Please enter in 'YYYY' format e.g. 1984
Q2	Q2 - What is your gender?
Q3	Q3 - Which state do you currently live in?
Q6ax1_1	Q6AX1 - In the last week, roughly how many hours did you spend listening to ANY music of your choosing? - Hour/s
Q6ax2_1	Q6AX2 - In the last week, roughly how many hours did you spend listening to ANY music of your choosing? - Minutes
Q19x1_1	Q19X1 - 1950_PlayASimpleMelody - Please click to play Have you heard this song before?
Q19_1_1	Q19 - 1950_PlayASimpleMelody - Please rate this piece of music.
Q19x1_2	Q19X1 - 1952_YouBelongToMe - Please click to play Have you heard this song before?
Q19_1_2	Q19 - 1952_YouBelongToMe - Please rate this piece of music.
Q19x1_3	Q19X1 - 1954_ShBoomSheBoom - Please click to play Have you heard this song before?
Q19_1_3	Q19 - 1954_ShBoomSheBoom - Please rate this piece of music.
Q19x1_4	Q19X1 - 1956_MyPrayer - Please click to play Have you heard this song before?
Q19_1_4	Q19 - 1956_MyPrayer - Please rate this piece of music.
Q19x1_5	Q19X1 - 1958_Patricia - Please click to play Have you heard this song before?
Q19_1_5	Q19 - 1958_Patricia - Please rate this piece of music.
Q19x1_6	Q19X1 - 1960_RunningBear - Please click to play Have you heard this song before?
Q19_1_6	Q19 - 1960_RunningBear - Please rate this piece of music.
Q19x1_7	Q19X1 - 1962_RosesAreRed - Please click to play Have you heard this song before?
Q19_1_7	Q19 - 1962_RosesAreRed - Please rate this piece of music.
Q19x1_8	Q19X1 - 1964_IGetAround - Please click to play Have you heard this song before?
Q19_1_8	Q19 - 1964_IGetAround - Please rate this piece of music.
Q19x1_9	Q19X1 - 1966_LastTrainToClarksville - Please click to play Have you heard this song before?
Q19_1_9	Q19 - 1966_LastTrainToClarksville - Please rate this piece of music.
Q19x1_10	Q19X1 - 1968_PeopleGotToBeFree - Please

We processed the dataset to produce two sets of data. The first data, `time_rating_data`, contains information on how long respondents listened to their chosen songs (`time_music_last_week`) and their average ratings (`rating_avg`). This data will later be used to conduct a correlation analysis between `time_music_last_week` and `rating_avg`. A summary of the first data set is presented in the Table 3.

```
# Data of time spend on music and rating
time_rating_data <- hw2_data |>
  select(time_music_last_week, Q19_avg) |>
  rename(rating_avg = Q19_avg) |>
  drop_na()

time_rating_data |>
  kbl() |>
  kable_styling(
    full_width = TRUE,
    bootstrap_options = c("striped", "condensed")
  )
```

The second data is `age_rating_data`. This data contains information on the age of the song (`song_age`) and its average rating (`rating_avg`). We will use this data to investigate the relationship between `song_age` and `rating_avg`. Table 4 shows the data.

```
# Data of age and song_age
age_data <- hw2_data |>
  select(Q1, starts_with("Q19_1_")) |>
  rename(birth_year = Q1) |>
  pivot_longer(
    cols = starts_with("Q19_1_"),
    names_to = "release",
    values_to = "rating"
  ) |>
  mutate(
    release = as.numeric(gsub("Q19_1_", "", release)) * 2 + 1948,
    song_age = release - birth_year
  )

# Data of song_age vs rating average
age_rating_data <- age_data |>
  select(song_age, rating) |>
  group_by(song_age) |>
  summarise(
```

Table 3: Respondent listening time and tatings data

time_music_last_week	rating_avg
2.4166667	8.2647059
2.0000000	4.9411765
5.5833333	7.6470588
15.0000000	5.8529412
24.0000000	3.9117647
4.0000000	2.2058824
6.3500000	7.4117647
6.5000000	3.1818182
10.0000000	5.0303030
5.0000000	5.6764706
21.4166667	8.8529412
21.4166667	8.5294118
2.0500000	3.8235294
6.0000000	7.1176471
14.0000000	7.5000000
3.9333333	9.5294118
8.0000000	1.2647059
6.0000000	6.2000000
8.2500000	9.8823529
1.0000000	9.5588235
15.0833333	2.2352941
2.4166667	9.4117647
5.5000000	6.5294118
5.5000000	5.6969697
1.0166667	9.5882353
12.5000000	8.8529412
1.5000000	5.0000000
3.0000000	6.0000000
10.0500000	0.6470588
2.0000000	6.5588235
2.5000000	9.1470588
2.0000000	7.8750000
7.0000000	4.2058824
6.5000000	9.6470588
4.0000000	6.1764706
12.0833333	7.6666667
5.7500000	9.7352941
3.0000000	5.0588235
4.0000000	7.7058824
2.5000000	6.5151515
10.5000000	5.6764706
5.5000000	9.4705882
4.5000000	8.8823529
2.1000000	2.5312500
10.0000000	5.8823529
5.0000000	6.5588235
5.0000000	4.6470588
4.0500000	8.1764706
2.0000000	8.2352941

```
    rating_avg = mean(rating, na.rm = TRUE),
    .groups = "drop"
  ) |>
  drop_na()

age_rating_data |>
  kbl() |>
  kable_styling(
    full_width = TRUE,
    bootstrap_options = c("striped", "condensed")
  )
```

Analysis

In Section , we will report our analysis results on the relationship between the duration of listening to chosen music and its ratings, as well as the relationship between the age of the song and its average ratings.

Time spend on music vs. rating

The relationship between the duration of listening to chosen music and the average ratings given can be observed using a scatter plot. Figure 1 presents this scatter plot.

```
time_rating_data |>
  ggplot(aes(x = time_music_last_week, y = rating_avg)) +
  geom_point(
    alpha = .4
  ) +
  geom_smooth(
    method = "lm",
    formula = y ~ x
  ) +
  theme_minimal()
```

Table 4: Song age and average ratings

song_age	rating_avg
-51	6.000000
-50	4.400000
-49	4.153846
-48	4.214286
-47	4.500000
-46	5.097561
-45	5.081967
-44	4.647059
-43	5.027778
-42	4.384615
-41	4.604938
-40	4.308642
-39	4.577320
-38	4.283019
-37	4.632479
-36	4.948718
-35	4.969466
-34	4.708029
-33	5.246835
-32	5.036364
-31	4.928994
-30	5.043478
-29	5.306011
-28	5.199005
-27	5.005076
-26	5.333333
-25	5.427230
-24	5.107296
-23	5.380342
-22	5.547325
-21	5.490272
-20	5.671533
-19	5.568266
-18	5.638796
-17	5.540351
-16	5.677019
-15	5.336667
-14	5.539394
-13	5.424051
-12	5.725146
-11	5.677914
-10	5.635328
-9	5.366864
-8	5.613079
-7	5.610028
-6	5.831579
-5	5.702186
-4	5.633588
-3	5.682170

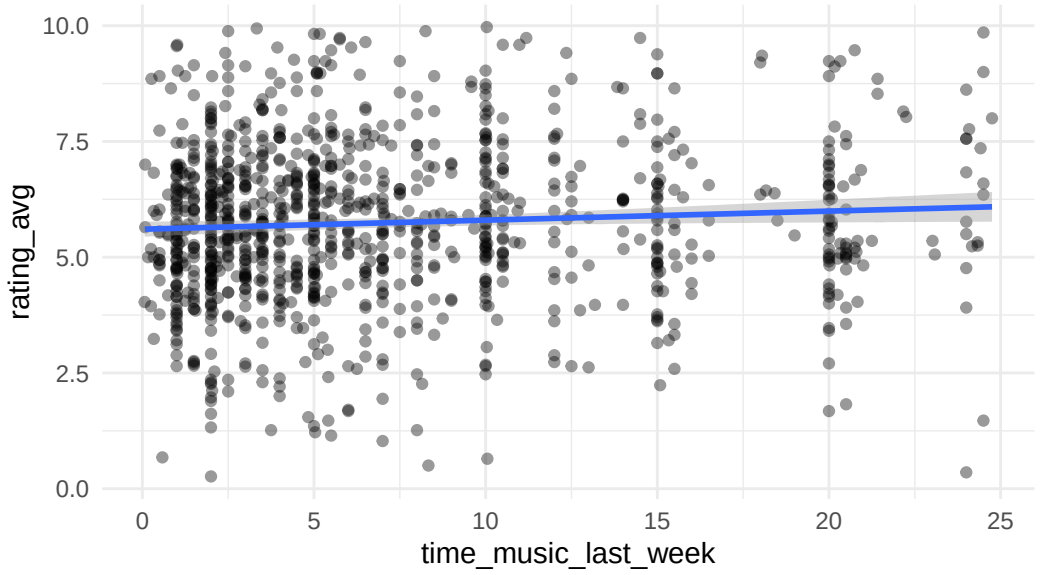

Figure 1: Scatter plot of listening duration vs. average ratings

To determine whether there is a correlation between the duration of listening to songs and the average ratings, a correlation hypothesis test needs to be conducted. The results are as follows.

```
cor_test_1 <- cor.test(
  x = time_rating_data$time_music_last_week,
  y = time_rating_data$rating_avg,
  method = "pearson"
)
print(cor_test_1)
```

```
    Pearson's product-moment correlation

data:  time_rating_data$time_music_last_week and time_rating_data$rating_avg
t = 2.2186, df = 1034, p-value = 0.02673
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.007959534 0.129196529
sample estimates:
       cor
0.06883216
```

Since we obtained a relatively large p-value, namely 0.0267291, we can conclude that there is not enough evidence to suggest a correlation.

Song age vs. rating average

Is there a relationship between the age of the song and its average rating? To observe this, first, please take a look at Figure 2.

```
age_rating_data |>
  ggplot(aes(x = song_age, rating_avg)) +
  geom_point() +
  geom_smooth(
    method = "lm",
    formula = y ~ x
  ) +
  theme_minimal()
```

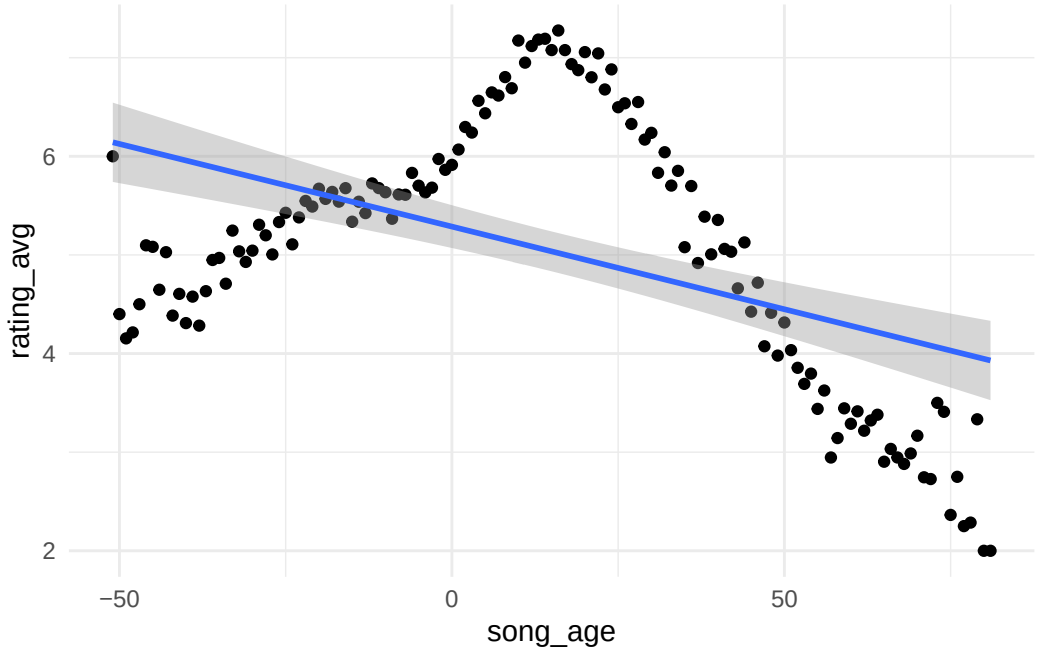

Figure 2: Scatter plot of song age vs. average ratings

To determine whether there is a correlation between the age of the song and its average rating, we need to conduct a correlation hypothesis test. Here are the results.

```r
cor.test(
  x = age_rating_data$song_age,
  y = age_rating_data$rating_avg,
  method = "pearson"
)
```

```
    Pearson's product-moment correlation

data:  age_rating_data$song_age and age_rating_data$rating_avg
t = -6.2962, df = 131, p-value = 4.254e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6027608 -0.3396278
sample estimates:
      cor
-0.481989
```

Since we obtained a very small p-value, there is enough evidence from the sample data to suggest that a correlation exists.

Is the correlation hypothesis test conducted appropriate? We performed a Pearson correlation hypothesis test on a non-linear relationship. This does not meet the assumptions of the hypothesis test. Therefore, we will analyze it further in the next section.

Song-age vs. rating average (revisited)

Although the data is not linear, we can split the data into two parts, resulting in two sections with a linear relationship. This division uses a song age threshold of 16.66 years. See Figure 3.

```r
age_rating_data_1 <- age_rating_data |>
  filter(song_age <= 16.66)
age_rating_data_2 <- age_rating_data |>
  filter(song_age > 16.66)
age_rating_data_group <- age_rating_data |>
  mutate(
    group = if_else(song_age <= 16.66, "lower", "upper")
  )
age_rating_data_group |>
  ggplot(aes(x = song_age, y = rating_avg, color = group)) +
```

```
  geom_point() +
  geom_smooth(
    method = "lm",
    formula = y ~ x
  ) +
  theme_minimal()
```

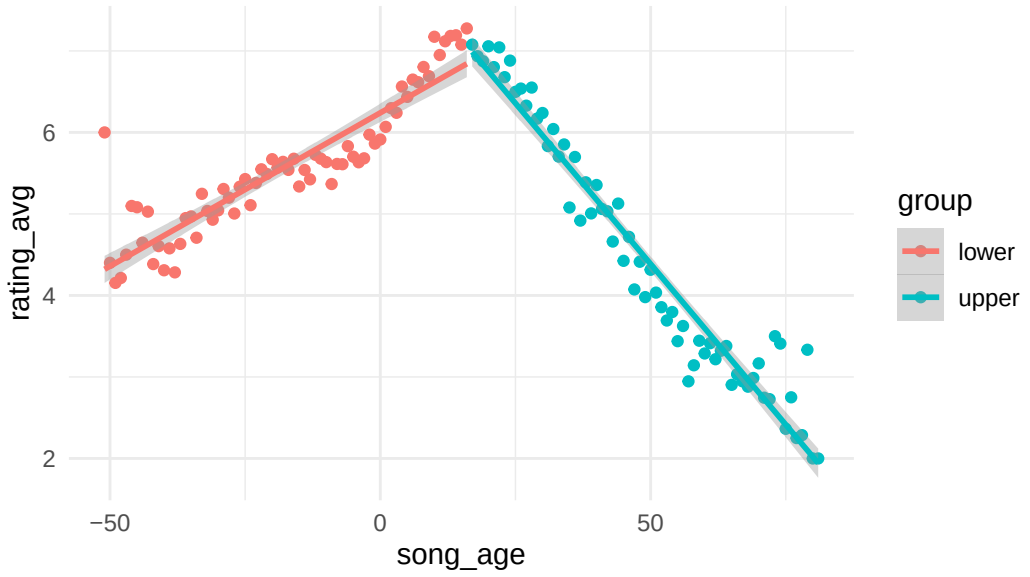

Figure 3: Scatter plot of song age vs. average ratings in grouped data

The results of the correlation hypothesis test for the first group are as follows. This test shows a significant correlation.

```
cor.test(
  x = age_rating_data_1$song_age,
  y = age_rating_data_1$rating_avg,
  method = "pearson"
)
```

```
    Pearson's product-moment correlation

data:  age_rating_data_1$song_age and age_rating_data_1$rating_avg
t = 17.664, df = 66, p-value < 2.2e-16
```

```
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8553808 0.9427346
sample estimates:
      cor
0.9085187
```

The second group also yields a significant correlation.

```
cor.test(
  x = age_rating_data_2$song_age,
  y = age_rating_data_2$rating_avg,
  method = "pearson"
)
```

```
    Pearson's product-moment correlation

data:  age_rating_data_2$song_age and age_rating_data_2$rating_avg
t = -32.611, df = 63, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9826619 -0.9537583
sample estimates:
       cor
-0.9716349
```

Key takeaways

We have conducted a correlation analysis on data related to music preferences. First, we found that there is not enough evidence from our sample data to suggest a correlation between the duration of listening to music and its average ratings.

Second, we found sufficient evidence to conclude that there is a correlation between the age of the song and its average rating. Furthermore, we showed that a song age of 16.66 years is associated with the highest average rating. This means that, generally, a person tends to prefer hit songs released around 17 years after their birth.