STA258 Spotify Songs Data Analysis - Group 10

Tidy Tuesday Project - Spotify

Yash Agarwal, Diaa Bakir, Angelo Gener, Steven Hua, Muhammad Iqbal

Research Question and Introduction to Step 3

We will be analyzing trends within a certain subset of the valence/danceability plot, along with their individual relationships with their year of album release, and comparing it through significance tests and p-values to the wider range of data to see if the smaller subset undermines the original data in any way. We selected the basis for this subset to be the year we were born (2001) as we wanted to assess whether the emotion in the songs we have listened to, was statistically different than the songs our parents/ teachers, etc. listened to. To analyze these trends, we would be analyzing three separate graphs looking at the release date vs danceability, release date vs valence, valence vs danceability to see whether there is correlation between the three plots based on potential spikes in the data points after certain years.

Do songs released in our lifetime have a greater impact on our emotions and our willingness to dance as compared to older songs released before 2001? In other words:

Question: Is there a strong correlation between how danceable a song is versus a song's valence rating after 2001?

Step 1: Specify the null and alternative hypotheses

Ho: Valence and Danceability are strongly correlated. Ha: There is no observable correlation between Valence and Danceability.

So we just take:

Ho: A = 0 Ha: A != 0

Step 2: State and check whether the assumptions about statistical model is met

Assumptions: We shall take the assumption that there does exist a strong correlation between dancability and valence in this particular subset of data. As such:

Valence and Danceability Subsets after 2001

We must first create the subset of songs that were released between 2001 and the present year.

```
track_album_release_date <- spotify_songs$track_album_release_date
valence <- spotify_songs$valence
danceability <- spotify_songs$danceability
track_album_release_date_new <- stringr::str_extract(track_album_release_date, "^.{4}")
track_album_release_date_year <- as.numeric(track_album_release_date_new)
spotify_songs_after_2001 <- subset(spotify_songs, track_album_release_date_year >= 2001)
```

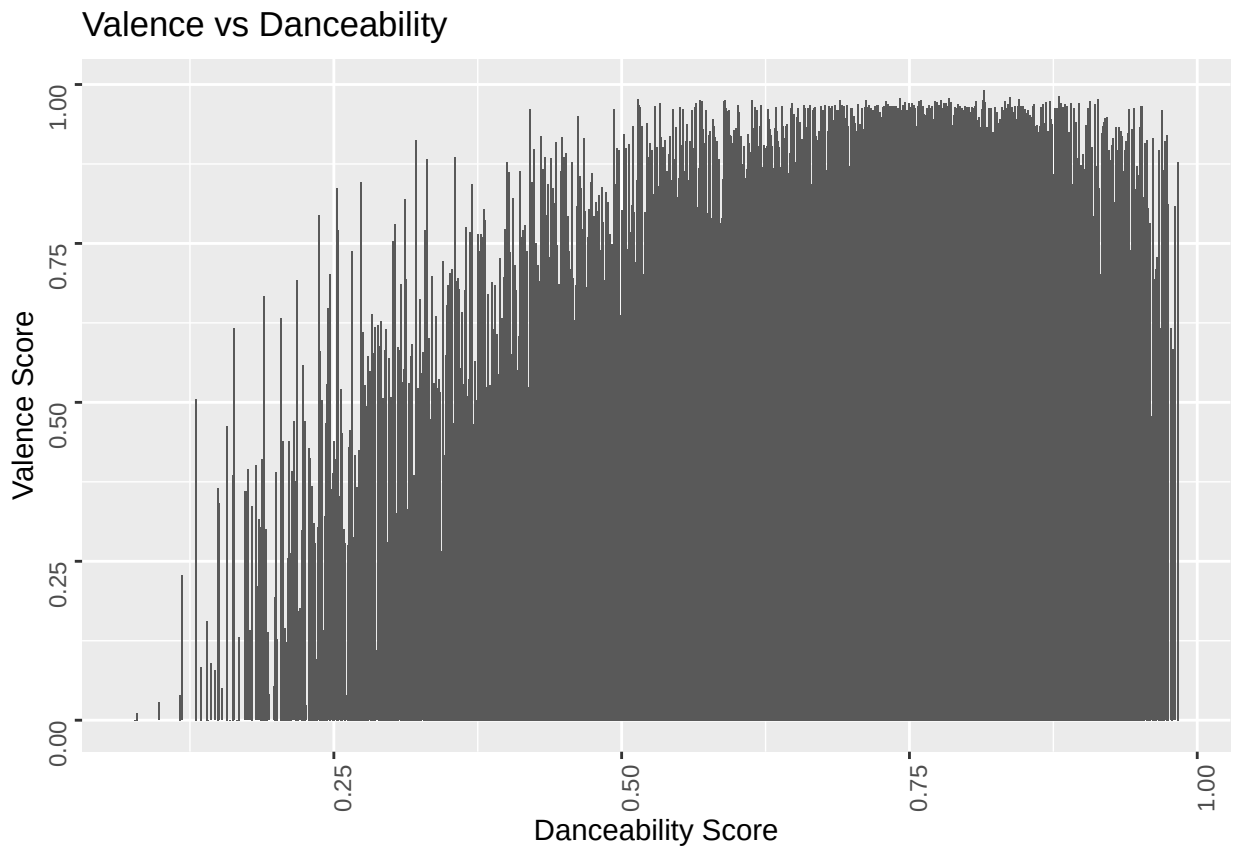
```
valence_after_2001 <- spotify_songs_after_2001$valence
favstats(valence_after_2001)
```

```
##    min    Q1 median    Q3   max      mean        sd     n missing
##  1e-05 0.313   0.49 0.672 0.991 0.4927269 0.2313529 27826       0
```

```
danceability_after_2001 <- spotify_songs_after_2001$danceability
favstats(danceability_after_2001)
```

```
##     min    Q1 median    Q3   max      mean        sd     n missing
##  0.0771 0.572  0.676 0.763 0.983 0.6607628 0.1404405 27826       0
```

```
ggplot(spotify_songs_after_2001, aes(y=valence_after_2001, x=danceability_after_2001)) +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text = element_text(angle=90, hjust=1)) +
  ggtitle("Valence vs Danceability") + xlab("Danceability Score") +
  ylab("Valence Score")
```



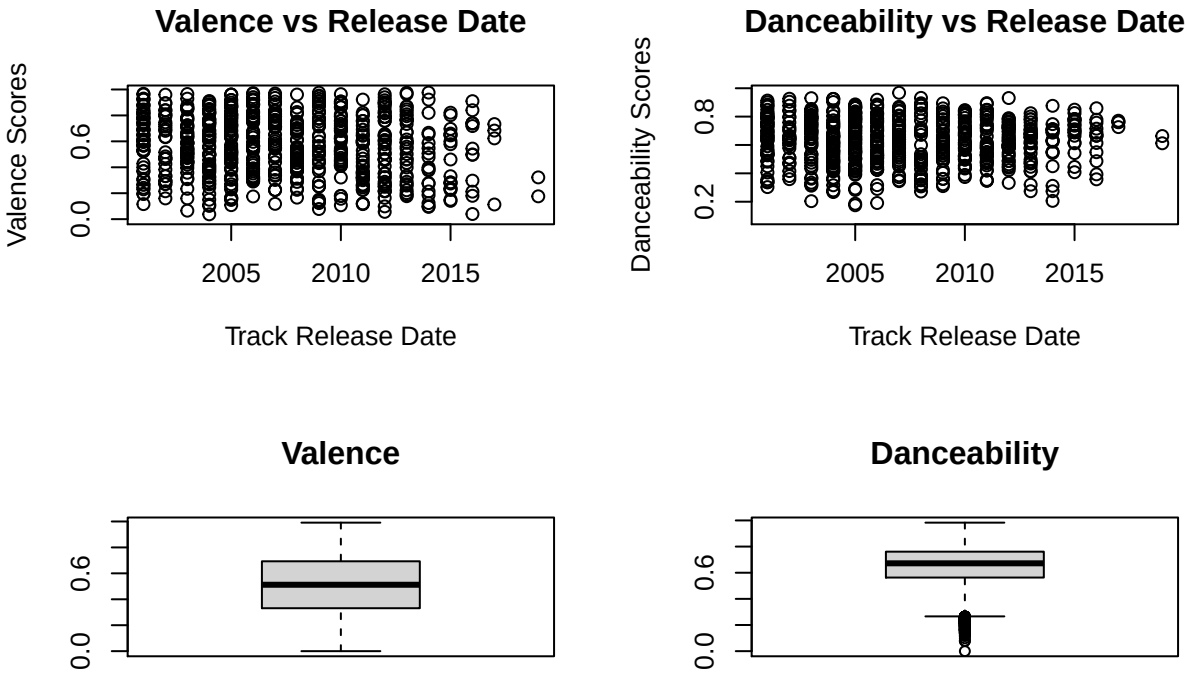

```
par(mfrow=c(2,2))
plot(spotify_songs_after_2001$track_album_release_date, valence_after_2001, main = "Valence vs Release
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion
```

```
plot(spotify_songs_after_2001$track_album_release_date, danceability_after_2001, main = "Danceability vs
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion
```

```
boxplot(valence, main = "Valence")
boxplot(danceability, main = "Danceability")
```



Step 3: State the value of the observed test-statistic

```
favstats(valence_after_2001)
```

```
##    min    Q1 median    Q3   max      mean        sd     n missing
##  1e-05 0.313   0.49 0.672 0.991 0.4927269 0.2313529 27826       0
```

```
favstats(danceability_after_2001)
```

```
##     min    Q1 median    Q3   max      mean        sd     n missing
##  0.0771 0.572  0.676 0.763 0.983 0.6607628 0.1404405 27826       0
```

```r
cov(valence_after_2001, danceability_after_2001)
```

```
## [1] 0.01107295
```

```r
cor(danceability_after_2001, valence_after_2001)
```

```
## [1] 0.3407973
```

```r
linear.model <- lm(valence_after_2001 ~ danceability_after_2001)
summary(linear.model)
```

```
##
## Call:
## lm(formula = valence_after_2001 ~ danceability_after_2001)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59174 -0.16500  0.00031  0.16363  0.61102
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.121769   0.006272   19.41   <2e-16 ***
## danceability_after_2001 0.561408   0.009285   60.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2175 on 27824 degrees of freedom
## Multiple R-squared:  0.1161, Adjusted R-squared:  0.1161
## F-statistic:  3656 on 1 and 27824 DF,  p-value: < 2.2e-16
```

Running the linear model function through R, we obtain a summary of the coefficients of Standard Error, t-values and the p-values.

t-value : 60.47 0.561408 / 0.009285

Step 4: State the p-value of the observed test-statistic

p-value: 2.2*10^-16

Step 5: Make a decision (e.g., reject Ho, fail to reject Ho) at the significane-level of = 0.05

As the p-value of the observed test-statistic is less than 0.05, we can conclude that we Reject Ho and conclude Ha.

Step 6: In plain, non-statistical language, give a conclusion from your analysis.

Obtaining the p-value of 2.2*10^-16 from our statistical t-value 67.54981, we can discern that the p-value is less than or equal to our significance level, as clearly...

= p-value <= significance-level = 2.2*10^-16 <= 0.05

The inequality holds true, hence, we are able to reject the null hypothesis as the data favours the alternative hypothesis (reject Ho and conclude Ha), indicating that our results are indeed statistically significant. In other words, we can conclude that there exists no strong observable correlation between valence and danceability in songs after 2001.