Main_project_file

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(ggplot2)
```

Importing data into R

```r
taylor_album_songs <- read_csv("/Users/donyabehroozi/Downloads/taylor_album_songs.csv")
```

```
Rows: 194 Columns: 29
-- Column specification --------------------------------------------------------
Delimiter: ","
chr   (7): album_name, track_name, artist, featuring, key_name, mode_name, k...
dbl  (14): track_number, danceability, energy, key, loudness, mode, speechin...
lgl   (4): ep, bonus_track, explicit, lyrics
date  (4): album_release, promotional_release, single_release, track_release

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
taylor_ratings <- read_csv("/Users/donyabehroozi/Desktop/stat365/stat-365/taylor_albums.cs
```

```
Rows: 14 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): album_name
dbl  (2): metacritic_score, user_score
lgl  (1): ep
date (1): album_release

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Merging data sets together

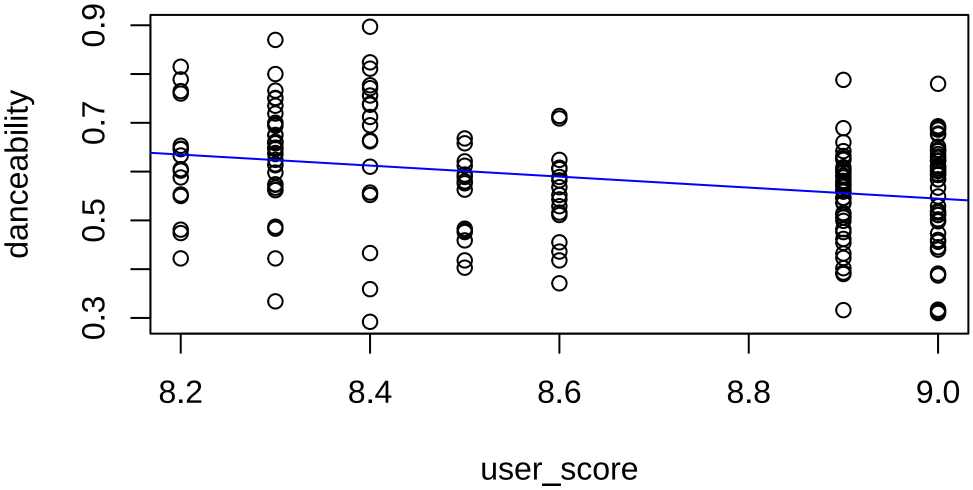
```
taylor_albums_joined <- taylor_album_songs |>
  left_join(taylor_ratings, by = c("album_name", ep = "ep", album_release = "album_release

head(taylor_albums_joined)
```

```
# A tibble: 6 x 31
  album_name   ep    album_release track_number track_name      artist featuring
  <chr>        <lgl> <date>               <dbl> <chr>           <chr>  <chr>    
1 Taylor Swift FALSE 2006-10-24               1 Tim McGraw      Taylo~ <NA>     
2 Taylor Swift FALSE 2006-10-24               2 Picture To Burn Taylo~ <NA>     
3 Taylor Swift FALSE 2006-10-24               3 Teardrops On M~ Taylo~ <NA>     
4 Taylor Swift FALSE 2006-10-24               4 A Place In Thi~ Taylo~ <NA>     
5 Taylor Swift FALSE 2006-10-24               5 Cold As You     Taylo~ <NA>     
6 Taylor Swift FALSE 2006-10-24               6 The Outside     Taylo~ <NA>     
# i 24 more variables: bonus_track <lgl>, promotional_release <date>,
#   single_release <date>, track_release <date>, danceability <dbl>,
#   energy <dbl>, key <dbl>, loudness <dbl>, mode <dbl>, speechiness <dbl>,
#   acousticness <dbl>, instrumentalness <dbl>, liveness <dbl>, valence <dbl>,
#   tempo <dbl>, time_signature <dbl>, duration_ms <dbl>, explicit <lgl>,
#   key_name <chr>, mode_name <chr>, key_mode <chr>, lyrics <lgl>,
#   metacritic_score <dbl>, user_score <dbl>
```

Linear regression analysis (danceability vs. user scores)

```
plot(danceability~user_score, data = taylor_albums_joined)
taylor.fit1=lm(danceability~user_score, data = taylor_albums_joined)
abline(taylor.fit1, col="blue")
```



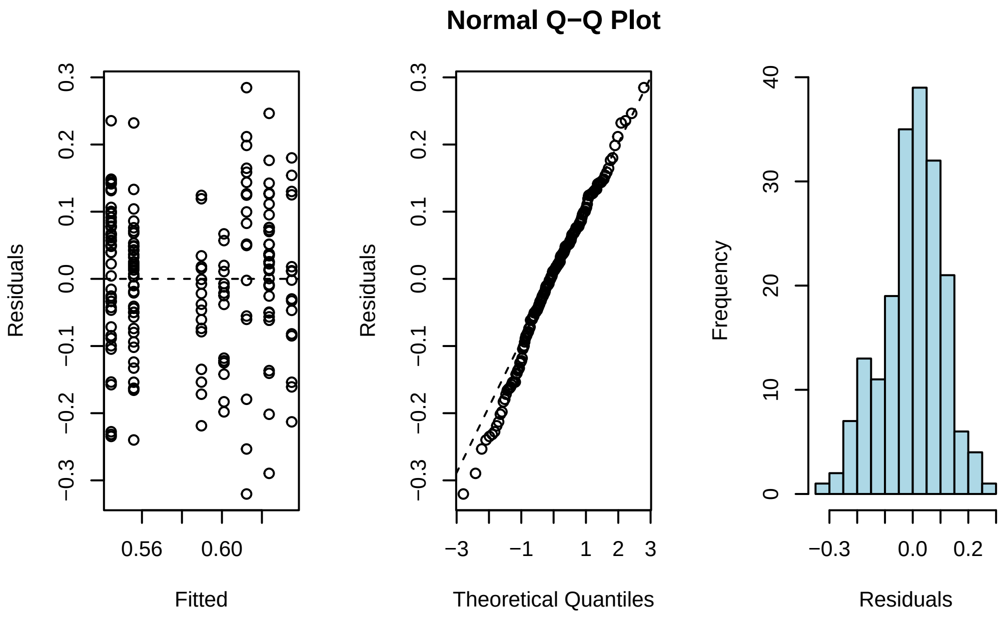
Hypotheses

H_0: Audience rating is not linearly associated with danceability score.

H_A: Audience rating is linearly associated with danceability score.

Checking simple linear regression assumptions

```
par(mfrow=c(1,3))
plot(resid(taylor.fit1)~fitted(taylor.fit1), xlab='Fitted', ylab='Residuals');abline(h=0,l
qqnorm(resid(taylor.fit1),ylab='Residuals'); qqline(resid(taylor.fit1),lty=2)
hist(resid(taylor.fit1),main="", xlab="Residuals",col='lightblue')
```



Results and conclusion

```
summary(taylor.fit1)
```

```
Call:
lm(formula = danceability ~ user_score, data = taylor_albums_joined)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32032 -0.06053  0.01211  0.07175  0.28468 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.56037    0.22873   6.822 1.18e-10 ***
user_score  -0.11286    0.02642  -4.271 3.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1105 on 189 degrees of freedom
```

```
  (3 observations deleted due to missingness)
Multiple R-squared:  0.08803,   Adjusted R-squared:  0.0832
F-statistic: 18.24 on 1 and 189 DF,  p-value: 3.076e-05
```

Given the very small p-value (close to 0) we have strong evidence to reject the null hypothesis. Therefore, we have sufficient evidence to claim that audience rating is linearly associated with danceability score for this population.

Linear regression analysis (danceability vs. metacritic scores)