MINI PROJECT 3 - DRAFT REPORT

Project Motivation :

I chose Taylor Swift data set from **https://github.com/rfordatascience/tidytuesday** whose source is **https://taylor.wjakethompson.com/** .

The recent Eras tour inspired me to explore this data set.

Research question :

To find the relationship between dancability and energy in the alubms of taylor swift.

Description of the table

This data was originally collected from genius and Spotify API

Main variables of interest :

album_name : Album name

dancebility : Spotify danceability score. A value of 0.0 is least danceable and 1.0 is most danceable.

energy : Spotify energy score. Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

loudness : Spotify loudness score. The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track.

Details of the table taylor_album_songs are as follows when Glimpse function is used

```
library(ggplot2)
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
```

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0
v readr     2.1.4

-- 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 becor
```

```r
taylor_album_songs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/t
```

```
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.
```

```
dplyr::glimpse(taylor_album_songs)
```

```
Rows: 194
Columns: 29
$ album_name          <chr> "Taylor Swift", "Taylor Swift", "Taylor Swift", "T~
$ ep                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
$ album_release       <date> 2006-10-24, 2006-10-24, 2006-10-24, 2006-10-24, 2~
$ track_number        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
$ track_name          <chr> "Tim McGraw", "Picture To Burn", "Teardrops On My ~
$ artist              <chr> "Taylor Swift", "Taylor Swift", "Taylor Swift", "T~
$ featuring           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ bonus_track         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
$ promotional_release <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ single_release      <date> 2006-06-19, 2008-02-03, 2007-02-19, NA, NA, NA, N~
$ track_release       <date> 2006-06-19, 2006-10-24, 2006-10-24, 2006-10-24, 2~
$ danceability        <dbl> 0.580, 0.658, 0.621, 0.576, 0.418, 0.589, 0.479, 0~
$ energy              <dbl> 0.491, 0.877, 0.417, 0.777, 0.482, 0.805, 0.578, 0~
$ key                 <dbl> 0, 7, 10, 9, 5, 5, 2, 8, 4, 2, 2, 8, 7, 4, 10, 5, ~
$ loudness            <dbl> -6.462, -2.098, -6.941, -2.881, -5.769, -4.055, -4~
$ mode                <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,~
$ speechiness         <dbl> 0.0251, 0.0323, 0.0231, 0.0324, 0.0266, 0.0293, 0.~
$ acousticness        <dbl> 0.57500, 0.17300, 0.28800, 0.05100, 0.21700, 0.004~
$ instrumentalness    <dbl> 0.00e+00, 0.00e+00, 0.00e+00, 0.00e+00, 0.00e+00, ~
$ liveness            <dbl> 0.1210, 0.0962, 0.1190, 0.3200, 0.1230, 0.2400, 0.~
$ valence             <dbl> 0.425, 0.821, 0.289, 0.428, 0.261, 0.591, 0.192, 0~
$ tempo               <dbl> 76.009, 105.586, 99.953, 115.028, 175.558, 112.982~
$ time_signature      <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,~
$ duration_ms         <dbl> 232107, 173067, 203040, 199200, 239013, 207107, 24~
$ explicit            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
$ key_name            <chr> "C", "G", "A#", "A", "F", "F", "D", "G#", "E", "D"~
$ mode_name           <chr> "major", "major", "major", "major", "major", "majo~
$ key_mode            <chr> "C major", "G major", "A# major", "A major", "F ma~
$ lyrics              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

```
head(taylor_album_songs)
```

```
# A tibble: 6 x 29
  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 22 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>
```

Structure of Data Frame is:

```
typeof(taylor_album_songs)
```

```
[1] "list"
```

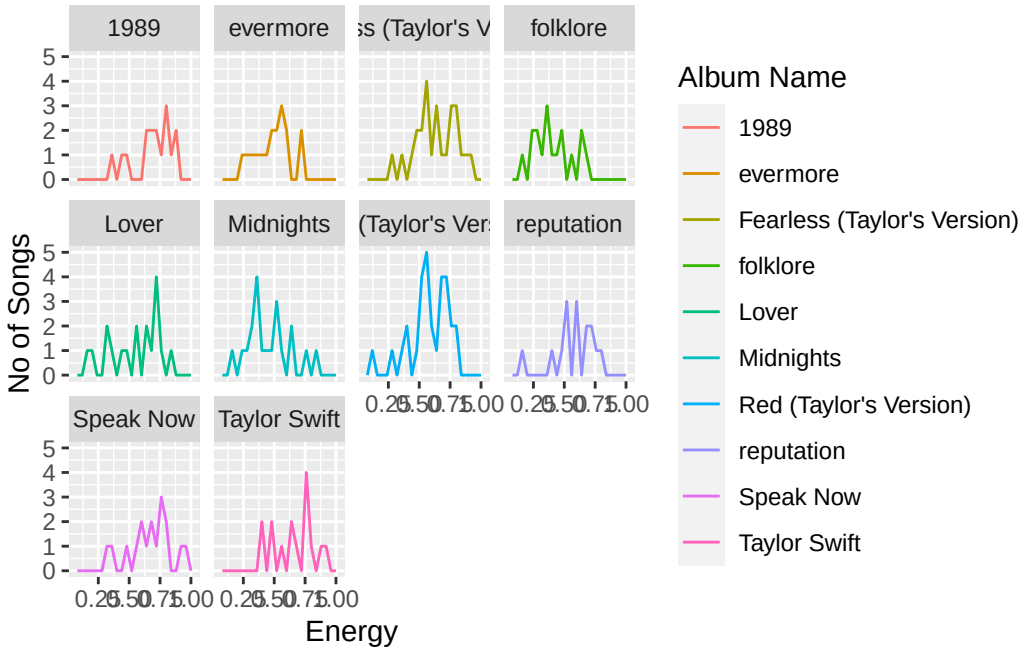
Exploratory data analysis -

PLOT 1:

Mapping of energy of the songs for each album

```
ggplot(taylor_album_songs, mapping=aes(energy, color=album_name)) +
  geom_freqpoly(binwidth=0.04) + facet_wrap(facets = vars(album_name))+
  scale_color_discrete(name="Album Name")+
labs(
       x = "Energy",
       y = "No of Songs",
    )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```



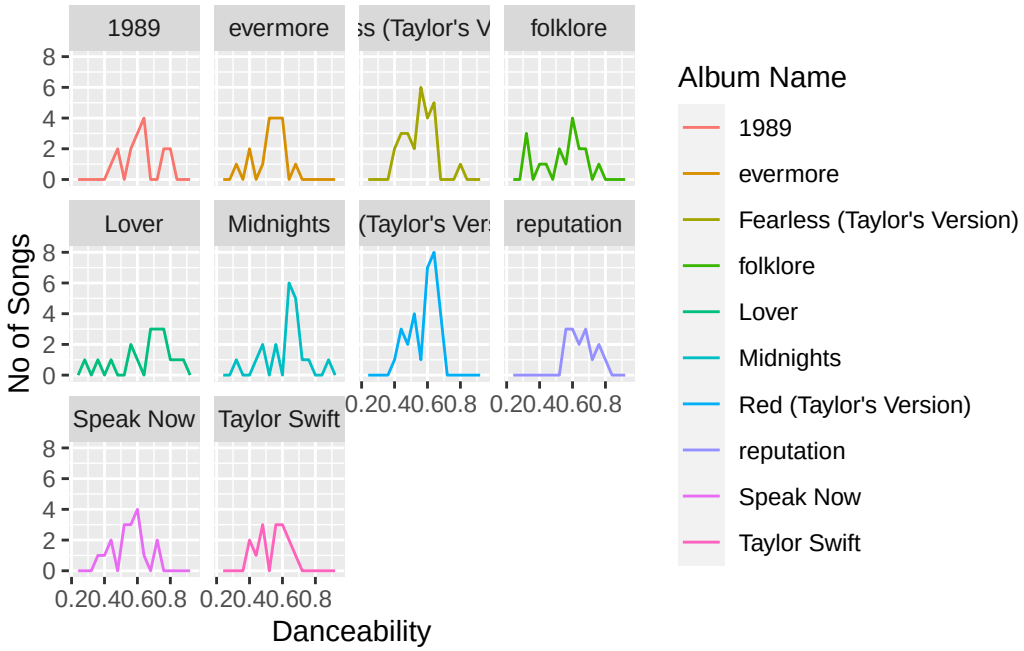
Summary : The songs with max energy are in the albums Red and Fearless Folklore has songs with lesser energy

PLOT 2:

Mapping of Danceability of songs for each album

```
ggplot(taylor_album_songs, mapping=aes(danceability, color=album_name)) +
  geom_freqpoly(binwidth=0.04) + facet_wrap(facets = vars(album_name))+
    scale_color_discrete(name="Album Name")+
labs(
       x = "Danceability",
       y = "No of Songs",
    )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```



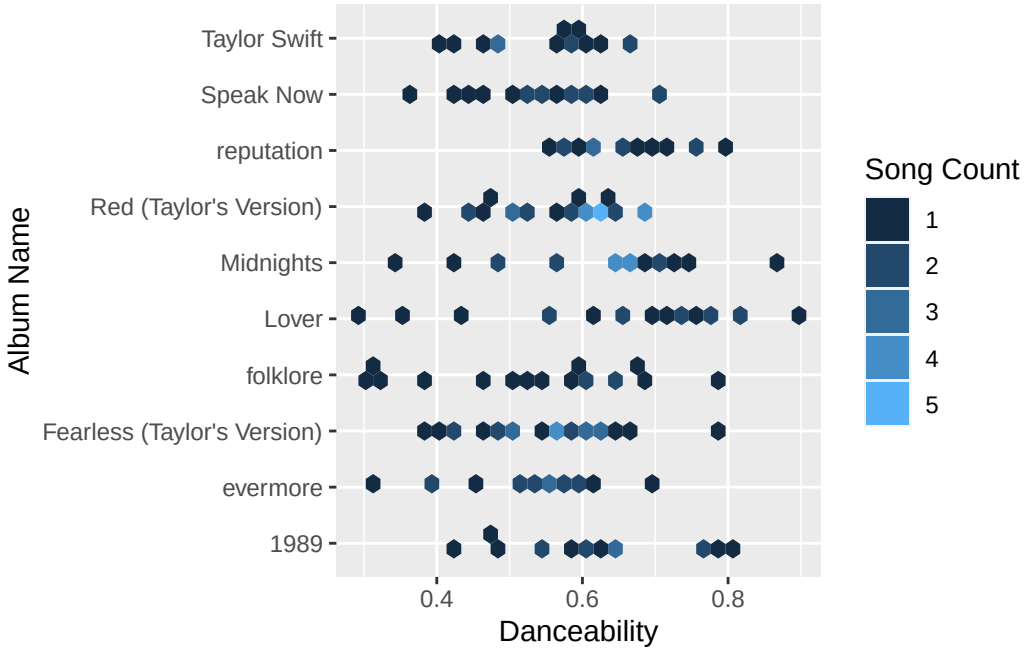
Summary : Reputation has majority of the songs in its album that are danceable.

PLOT 3:

Mapping of Danceability of each song in every album, each hexagon represents a song. Song count is no of songs in the same danceability bucket

```
ggplot(taylor_album_songs, mapping=aes(x=danceability, y=album_name)) +
  geom_hex() +
  guides(fill=guide_legend(title="Song Count"))+
labs(
       x = "Danceability",
       y = "Album Name",
    )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_binhex()`).
```



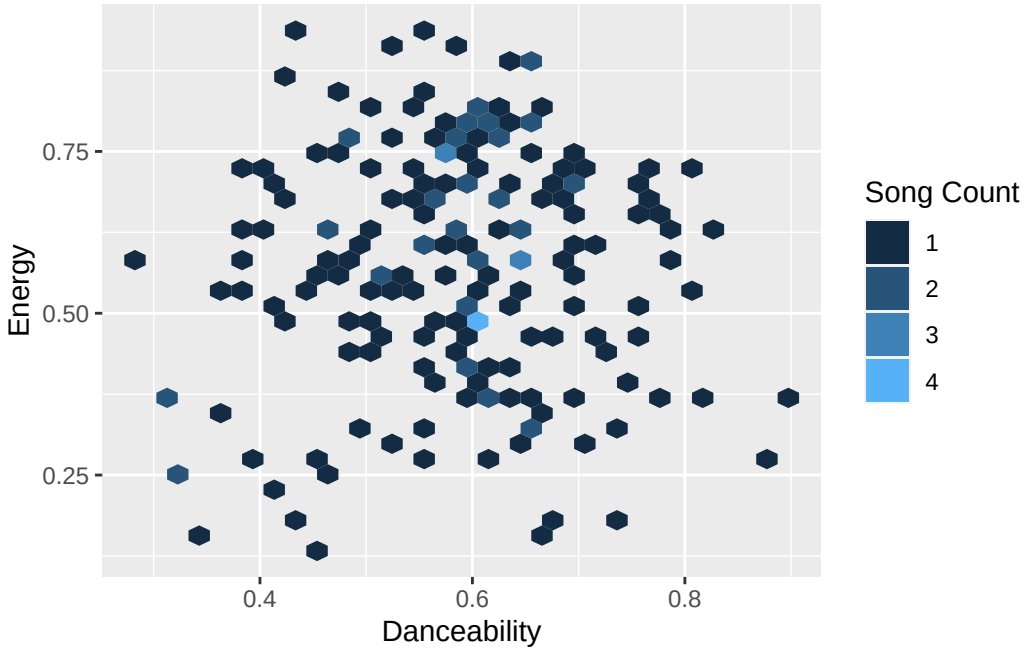
Summary : Danceability of songs in the album Red is more grouped around 0.6 but in the album Lover the songs have danceability thats spread wide ranging from 0.2 to 1.0

PLOT 4:

Mapping of dancebility and energy

```
ggplot(taylor_album_songs, mapping=aes(x=danceability, y=energy)) +
  geom_hex()+
  guides(fill=guide_legend(title="Song Count"))+
labs(
       x = "Danceability",
       y = "Energy",
     )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_binhex()`).
```



Summary : Danceability and energy of most songs is in and around the point of intersection of 0.6 and 0.75

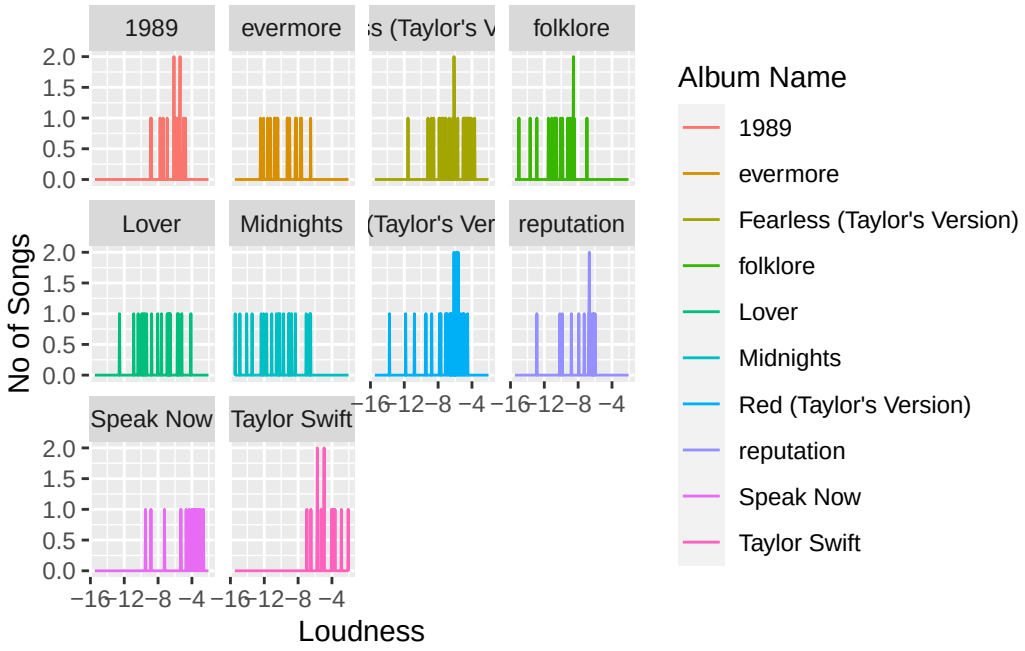
Danceability and Energy against other variables and their patterns

PLOT 5:

Mapping loudness for all songs per album

```
ggplot(taylor_album_songs, mapping=aes(loudness, color=album_name)) +
  geom_freqpoly(binwidth=0.04) + facet_wrap(facets = vars(album_name))+
  scale_color_discrete(name="Album Name")+
labs(
       x = "Loudness",
       y = "No of Songs",
    )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```



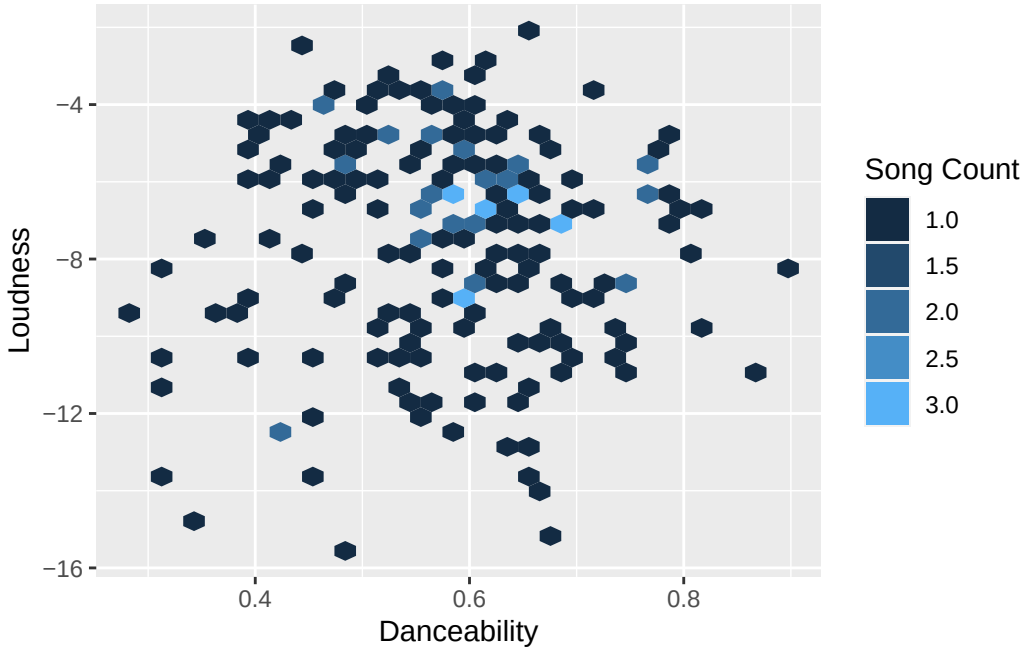
Summary : The values for loudness doesnt look continuous and are spread out. Max 2 songs are in one bucket of loudness Midnight has lot of songs with high loudness, Speak Now album has low loudness

PLOT 6:

Mapping danceability vs loudness for all songs, each hexagon is a song, loudness is negative as it is in decibels.

```
ggplot(taylor_album_songs, mapping=aes(x=danceability, y=loudness)) + geom_hex()+
guides(fill=guide_legend(title="Song Count"))+
labs(
       x = "Danceability",
       y = "Loudness",
     )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_binhex()`).
```



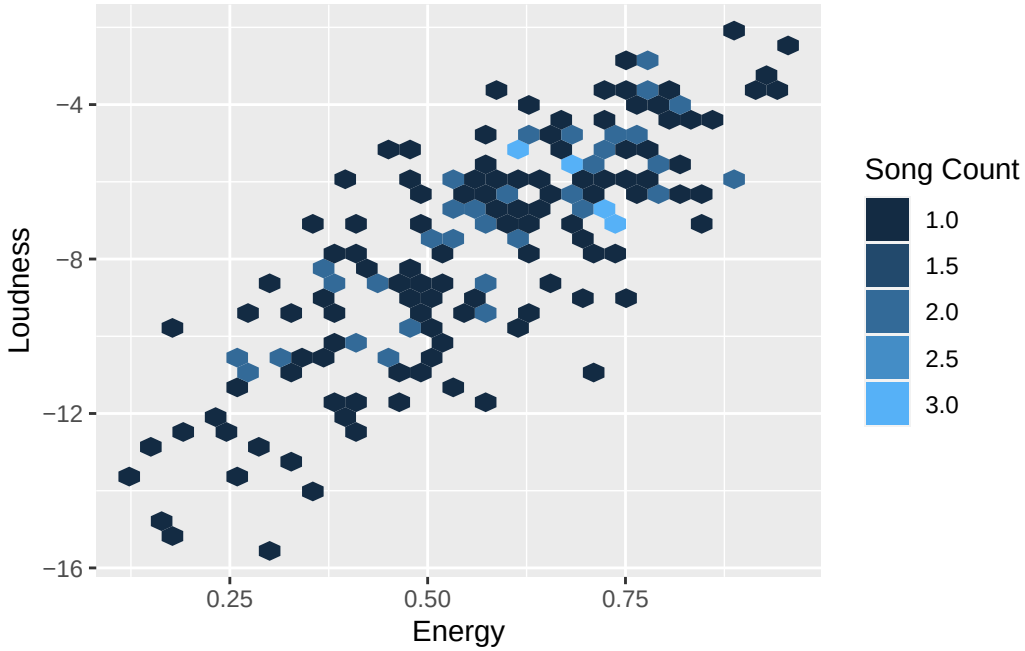
Summary : Most of Taylor swift songs are medium danceable(around 0.6) and less loud (between -4dB and -8dB).

PLOT 7:

Mapping energy vs loudness for all songs, each hexagon is a song, loudness is negative as it is in decibels.

```
ggplot(taylor_album_songs, mapping=aes(x=energy, y=loudness)) + geom_hex()+
  guides(fill=guide_legend(title="Song Count"))+
labs(
       x = "Energy",
       y = "Loudness",
     )
```

```
Warning: Removed 3 rows containing non-finite values (`stat_binhex()`).
```



Summary : Most of taylors songs are high on energy(0.6-0.8) and low on loudness(-4dB to -8dB) Energy and loudness appear to be highly correlated.

Correlation Tables

Co-relation between Danceability and energy

```
taylor_album_songs = taylor_album_songs %>%
  drop_na(danceability) %>%
  drop_na(energy)

taylor_album_songs %>%
  #filter(is.na(danceability,energy)=FALSE) %>%
  group_by(album_name) %>%
  summarize(correlation = cor(danceability, energy),songs_ct= n(),mean_dan=mean(danceabili
```

```
# A tibble: 10 x 7
   album_name                  correlation songs_ct mean_dan mean_en sd_dan sd_en
   <chr>                             <dbl>    <int>    <dbl>   <dbl>  <dbl> <dbl>
 1 1989                            -0.0312       16    0.624   0.697 0.116  0.153
 2 Fearless (Taylor's Versio~       0.0254       26    0.551   0.639 0.0915 0.161
```

```
 3 Lover                          0.0527       18    0.658    0.545 0.164  0.202
 4 Midnights                      0.125        20    0.627    0.451 0.122  0.169
 5 Red (Taylor's Version)         0.179        30    0.577    0.587 0.0821 0.157
 6 Speak Now                     -0.182        17    0.549    0.667 0.0941 0.178
 7 Taylor Swift                   0.0413       15    0.545    0.664 0.0852 0.166
 8 evermore                       0.487        17    0.527    0.494 0.0921 0.140
 9 folklore                       0.199        17    0.542    0.416 0.142  0.156
10 reputation                    -0.300        15    0.658    0.583 0.0751 0.163
```

Observations : In general the correlation between danceability and energy isnt prominent. Evermore with a value close to 0.5 has highest corrlation between danceability and energy. And, as deduced from the plots earlier Lover has more spreadout in danceability with maximum Standard Deviation and Red has the least spread in danceability.

Co-relation between Energy and Loudness

```
taylor_album_songs = taylor_album_songs %>%
  drop_na(energy) %>%
  drop_na(loudness)

taylor_album_songs %>%
  #filter(is.na(energy,loudness)=FALSE) %>%
  group_by(album_name) %>%
  summarize(correlation = cor(energy, loudness),n= n(),mean_en=mean(energy,na.rm=TRUE),mea
```

```
# A tibble: 10 x 7
   album_name                  correlation     n mean_en mean_loud sd_en sd_loud
   <chr>                             <dbl> <int>   <dbl>     <dbl> <dbl>   <dbl>
 1 1989                              0.654    16   0.697     -6.32 0.153    1.27
 2 Fearless (Taylor's Version)       0.749    26   0.639     -6.20 0.161    1.89
 3 Lover                             0.844    18   0.545     -8.01 0.202    2.26
 4 Midnights                         0.808    20   0.451    -10.6  0.169    2.60
 5 Red (Taylor's Version)            0.840    30   0.587     -7.01 0.157    2.10
 6 Speak Now                         0.884    17   0.667     -4.66 0.178    2.03
 7 Taylor Swift                      0.839    15   0.664     -4.73 0.166    1.34
 8 evermore                          0.520    17   0.494     -9.78 0.140    1.71
 9 folklore                          0.620    17   0.416    -10.3  0.156    2.08
10 reputation                        0.861    15   0.583     -7.67 0.163    1.95
```

Observations : Energy and loudness are highly correlated with correlation coefficient being around 0.8 for most albums. Almost all albums have similar standard deviation for energy

meaning the pattern of spread of energy for songs is similar in all albums. Lover album is one exception to this.

Co-relation between Danceability and Loudness

```
taylor_album_songs = taylor_album_songs %>%
  drop_na(danceability) %>%
  drop_na(loudness)

taylor_album_songs %>%
  #filter(is.na(Danceability,loudness)=FALSE) %>%
  group_by(album_name) %>%
  summarize(correlation = cor(danceability, loudness),n= n(),mean_dan=mean(danceability,na
```

```
# A tibble: 10 x 7
   album_name                correlation     n mean_dan mean_loud sd_dan sd_loud
   <chr>                           <dbl> <int>    <dbl>     <dbl>  <dbl>   <dbl>
 1 1989                          0.362      16    0.624     -6.32 0.116     1.27
 2 Fearless (Taylor's Versi~    -0.171      26    0.551     -6.20 0.0915    1.89
 3 Lover                         0.209      18    0.658     -8.01 0.164     2.26
 4 Midnights                     0.327      20    0.627    -10.6  0.122     2.60
 5 Red (Taylor's Version)       -0.00239    30    0.577     -7.01 0.0821    2.10
 6 Speak Now                    -0.0153     17    0.549     -4.66 0.0941    2.03
 7 Taylor Swift                  0.137      15    0.545     -4.73 0.0852    1.34
 8 evermore                      0.170      17    0.527     -9.78 0.0921    1.71
 9 folklore                      0.165      17    0.542    -10.3  0.142     2.08
10 reputation                   -0.133      15    0.658     -7.67 0.0751    1.95
```

Observations : Correlation between Danceability and loudness is very low. Most of the albums have SD around the value 2 or closer to 2. Most albums have similar spread for loudness.

Methodology

Description of Analysis Process

Analysed 3 prominent variables from the data set taylo_album_songs - energy, danceability and loudness. Plotted the histograms for each of the variables with respect to individual album. Compared dependency of one variable over the other using correlation plots to derive a relation between them. Tried to look at the same using correlation coefficients in each album along with mean and standard deviation of each variable.

Reason for chossing this Analysis Process for the research question and how you arrived to final answer

- The reason why histogram was chosen is to get the distribution of each of these variables in each album. Histogram is a better visual aid to look at the frequency of songs over the variables.
- Correlation graphs were used to see the spread of variables and their dependency on each other. The points in the graph appear closer to a straight line if the variables are strongly correlated.
- The summary tables will give a better estimate of dependency with numeric values to look at. While the correlation plots are helpful for visualisation, the tables are for calculations.

Answer to research question in 2 paragraphs

From the analysis we can see that energy and loudness are related across albums. High energy songs are very loud. But Danceability isnt strongly related to either of these variables. Most of the songs have a danceability of 0.6 , meaning they are neither high nor too low on danceability. But most of the songs are high on energy(0.6-0.8) and low on loudness(-4dB to-8dB) even though they are strongly correlated. That is because overall loudness iself is low.

Looking at each album individually, - Album Red has maximum energy while album Fearless has less energy - Danceability of songs in the album Red is more grouped but in the album Lover the songs have danceability thats spread wide. - Midnight has lot of songs with high loudness, Speak Now album has low loudness