Impact of Tempo and Valence on Taylor Swift's Discography - A Case Study

Mithun Meenakshi S

2024-04-13

About This Case Study:

This case study was done as a part of Google's Data Analytics Certification Programme, which offered the option of pursuing an independent case study on any topic. This case study was started on 13-04-2024, with only a week to go for the release of Taylor Swift's 11th studio album, "The Tortured Poets Department". Beinga fan of Swift for almost 3 years, I wanted to do something to commemorate this release, and this case study is an attempt at the same. The data set contains data from Spotify about all her songs available to stream and the data set can be found here. Hope you enjoy this analysis!

Data Manipulation:

The original dataset, named "taylor_swift_spotify.csv" contains details like id, release_date, spotify url et cetera, but I have made a copy and removed data about the id and url using Microsoft Excel. Furthermore, I have extracted the years from the release_date column to release_year column, and removed release_dates from the working file.

Business question:

The main objective of this case study is to determine whether a song's tempo and valence impact the overall popuularity of a song in Taylor Swift's discography. This analysis was conducted mainly using Rstudio cloud, and I am planning to create a Tableau dashboard soon, which will be linked later.

Programming and Analysis:

The tidyverse package was mainly used to perform this analysis and the installation and loading was done using the following code:

```
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
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   1.5.1
## v ggplot2   3.5.0     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 become error
```

The tidyverse package contains many useful subpackages such as tidyr, ggplot2, dplyr, readr et cetera. In this analysis we have primarily used readr and ggplot2. The working file which was modified from the source file in earlier steps was opened and then examined using readr as follows:

```
taylor_df <- read_csv("working file.csv")
```

```
## New names:
## Rows: 530 Columns: 17
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (3): name, album, release_date dbl (14): ...1, release_year, track_number,
## acousticness, danceability, ener...
## 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.
## * `` -> `...1`
```

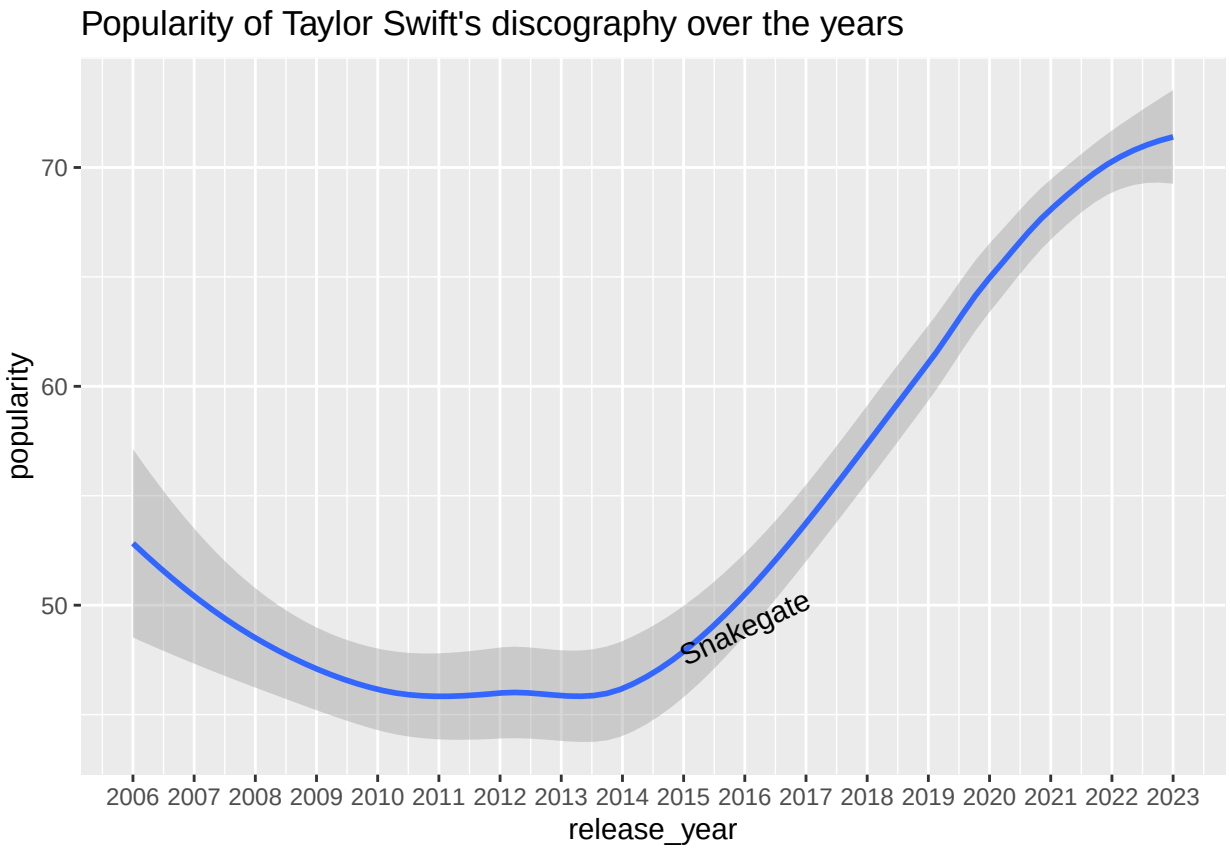
```
head(taylor_df)
```

```
## # A tibble: 6 x 17
##    ...1 name           album release_date release_year track_number acousticness
##   <dbl> <chr>          <chr> <chr>               <dbl>        <dbl>        <dbl>
## 1     0 Welcome To Ne~ 1989~ 10/27/2023           2023            1     0.00942
## 2     1 Blank Space (~ 1989~ 10/27/2023           2023            2     0.0885
## 3     2 Style (Taylor~ 1989~ 10/27/2023           2023            3     0.000421
## 4     3 Out Of The Wo~ 1989~ 10/27/2023           2023            4     0.000537
## 5     4 All You Had T~ 1989~ 10/27/2023           2023            5     0.000656
## 6     5 Shake It Off ~ 1989~ 10/27/2023           2023            6     0.0121
## # i 10 more variables: danceability <dbl>, energy <dbl>,
## #   instrumentalness <dbl>, liveness <dbl>, loudness <dbl>, speechiness <dbl>,
## #   tempo <dbl>, valence <dbl>, popularity <dbl>, duration_ms <dbl>
```

The parameters for various aspects of a song contained in this dataset are clearly explained on the Kaggle website of the dataset, which can be found here. Our analysis primarily focuses on popularity, acousticness, valence, and tempo. Our first discussion is about the popularity of Swift's songs over the years, which was visualised using ggplot2 package as follows:

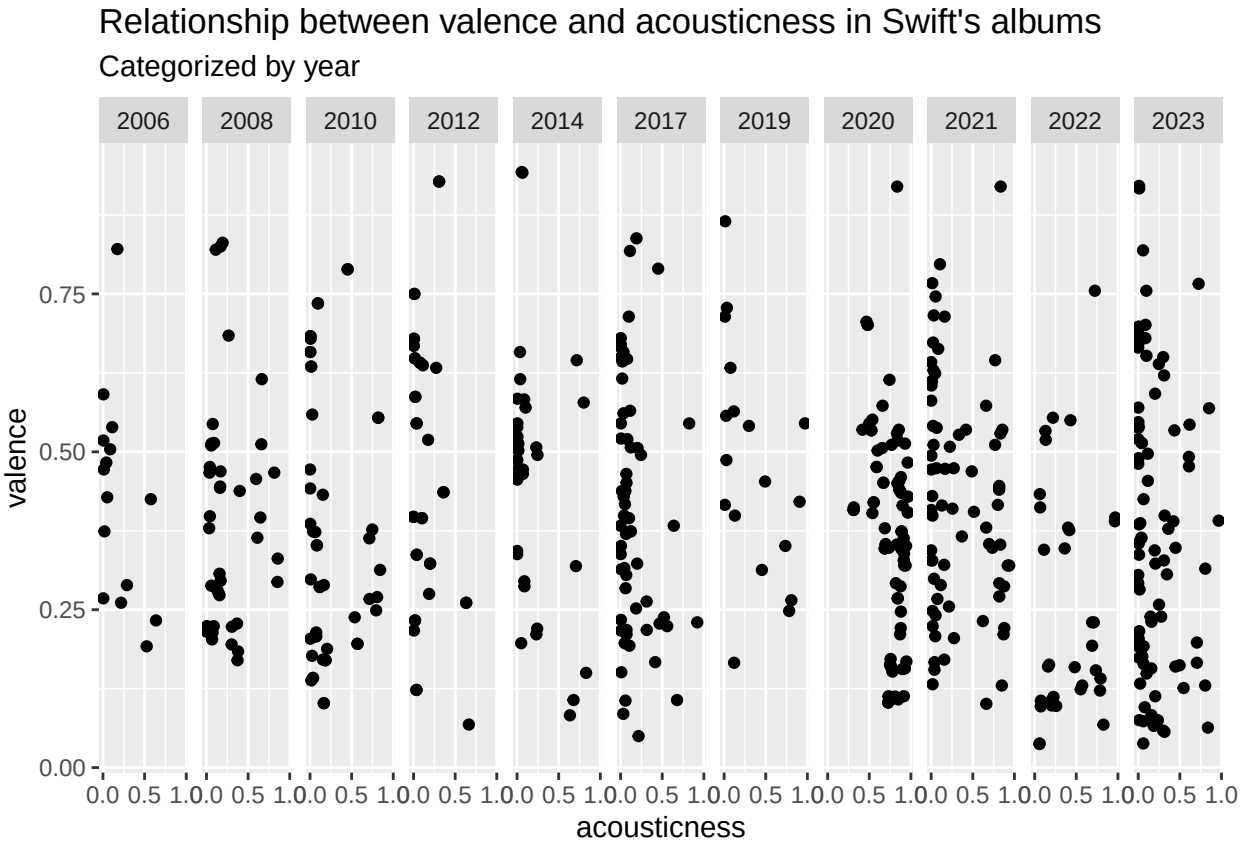
```
ggplot(data = taylor_df) +  geom_smooth(mapping = aes(x = release_year, y = popularity)) + labs(title =
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Here "Snakegate" refers to the feud between Swift and Kanye West, which subsequently led to the release of "Look What You Made Me Do" and skyrocketed her popularity. Next I focused on the relationship between acousticness and valence, categorised by years, which was accomplished using the code below:

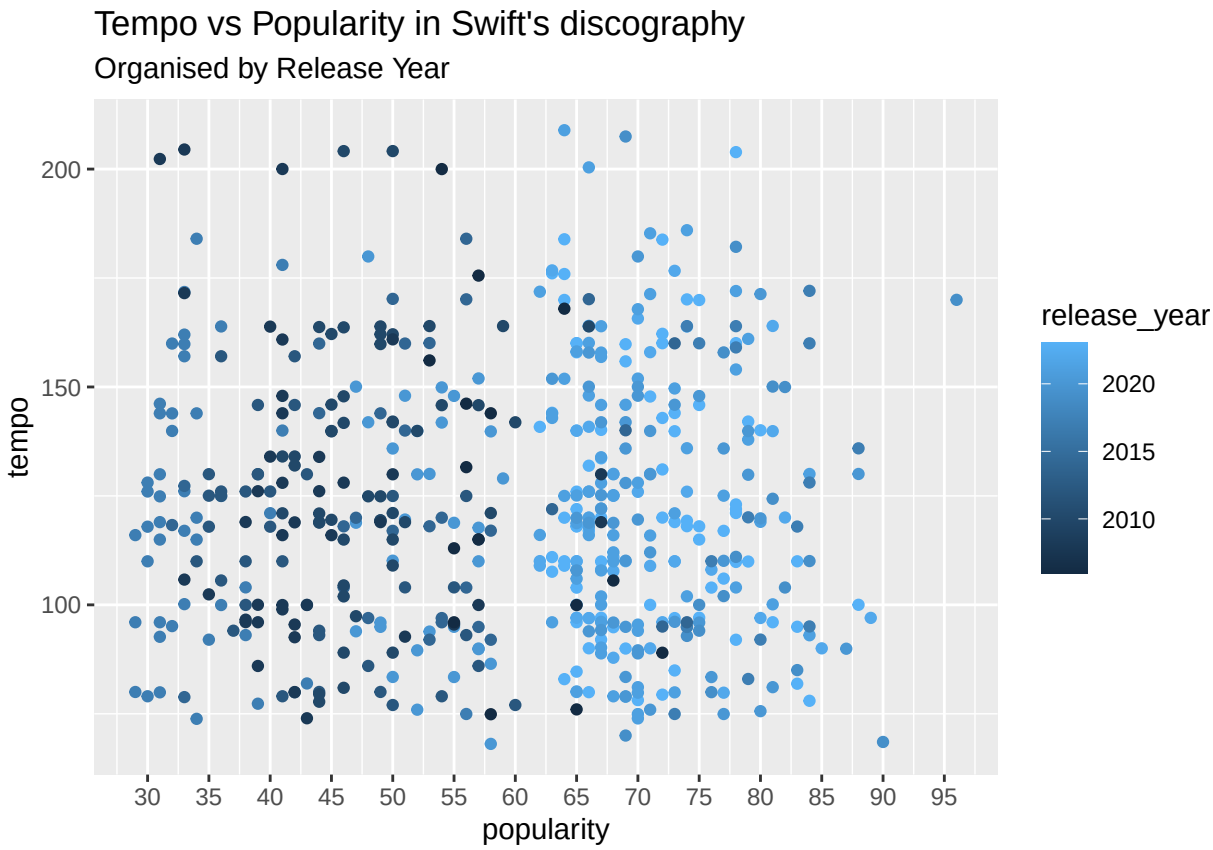
```
ggplot(data = taylor_df) + geom_point(mapping = aes(x = acousticness, y = valence)) + facet_grid(~releas
```



This data visualisation is pretty interesting, as we can understand the overall sound and mood of the albums released in the particular years. For example, we can see that most of her albums heavily lean towards the left side of the acousticness scale, being least acoustic, except for one. In 2020, Swift released a suprise album titled 'Folklore', which is a divergence from both Country and Pop. This album is heavily acoustic and the graph supports this notion. Furthermore, there is a misconception among the general public that Taylor Swift only writes breakup songs or songs about her exes. This is completely wrong, and this graph supports that. We can see that most of her albums lean towards the median valence (valence describes the overall mood of the song) disproving the general public's notion.

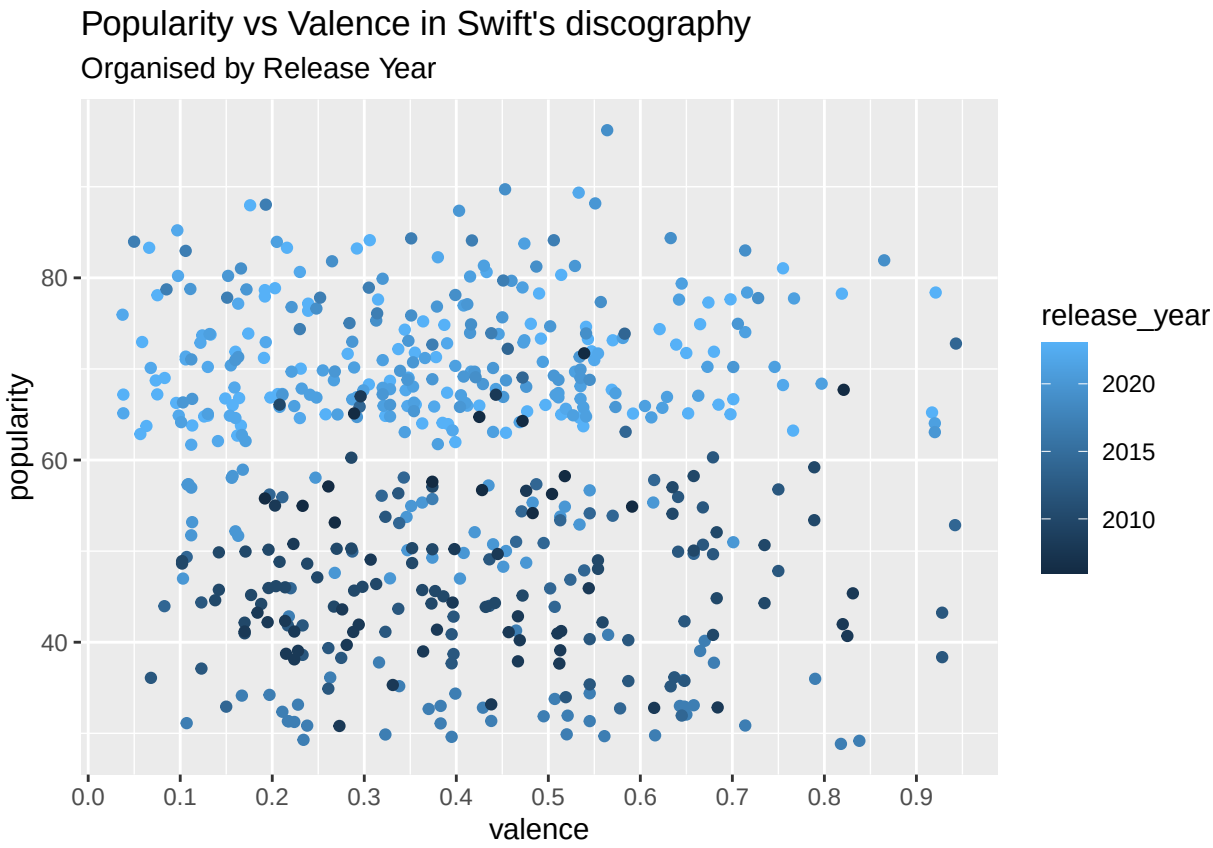
"Shake It Off" is a global hit in Swift's discography due to the overall danceability and tempo, but can we say the same to the rest of Swift's discography? Do songs have to have fast tempo to be global hits? Let us explore the relationship between tempo and popularity of songs in Swift's discography as follows:

```
ggplot(data = taylor_df) + geom_point(mapping = aes(x = popularity, y = tempo, colour = release_year)) +
```



We can immediately understand that the above questions are answered by this visualisation. There are quite a few data points with tempo greater than 200 but rank less than 50 on the popularity scale. There is one song with low tempo but ranking at 90 on the popularity scale (Lover). This visualisation doesn't reveal a straightforward relationship between tempo and popularity in Swift's discography. Let us explore the relationship between valence and popularity in Swift's discography.

```
ggplot(data = taylor_df) + geom_jitter(mapping = aes(x = valence, y = popularity, colour = release_year
```



Swift's discography contains songs with a wide range of emotions, and the data points scattered across the valence scale is proof to that. Most of Swift's songs lie between a valence of 0.1 (sad) to 0.7 (cheerful), and this is a testament to the varied themes of her discography

Conclusion:

This analysis gives us an overview of the moods and nature of Taylor Swift's discography, and proves that valence and tempo do not impact the popularity. I hope you enjoyed this analysis, and I will link a Tableau dashboard soon.