# What makes a song “classic”: an analysis of lyrics from top songs from 1960 to 2005?

Kuan-Jung Huang[1], Jiacheng Wang[2], Zhangqi Duan[3]

[1] Psychological and Brain Sciences
[2] Chemical Engineering
[3] Computer Science

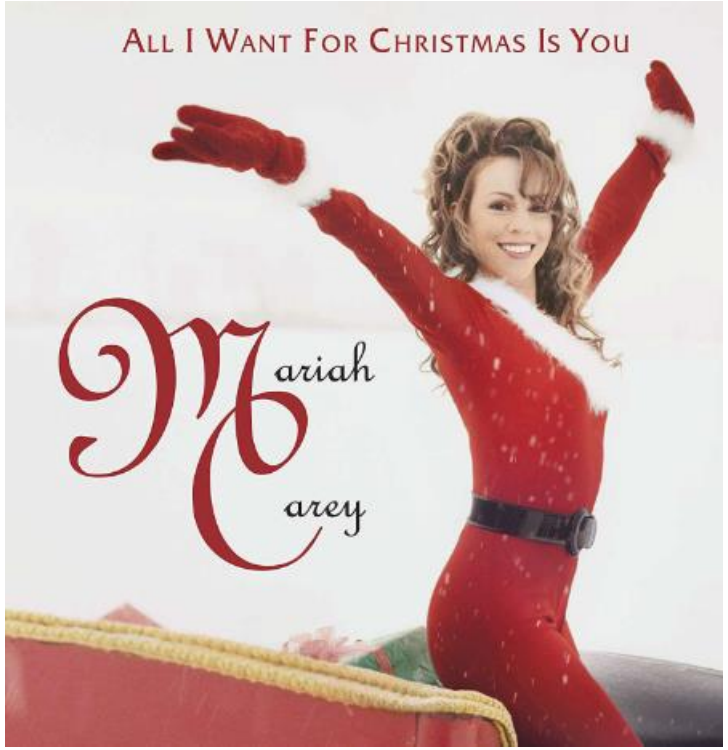
STATS 535 Group 2 Final Project Presentation

# Outline

- Introduction (Research Question & Dataset)
- Textual processing/analysis
- Webscraping
- Nonparametric analysis
- Power analysis
- Parametric analysis
- Conclusion

# Background (motivation)


College Music Classes on Taylor Swift & Other Pop Stars



# Research Question

- Does the times of repetition of the song title explain how popular a song is?

- Hypothesis:
  the song title repeats ↑ memorability & catchiness ↑, hence popularity ↑

- Popularity (YouTube Viewcount) ~ song title repetition times

# Original dataset

- https://www.kaggle.com/datasets/stefancomanita/top-us-songs-from-1950-to-2019-w-lyrics
- 5 columns and 700 rows (10 top songs per year, for 70 years)
- Lines of lyrics are segmented with |

| | year | rank | artist | song | lyrics |
|---|---|---|---|---|---|
| 1 | 1960 | 1 | Shirelles | Will You Love Me Tomorrow | Tonight you're mine completely\|You give your love so sweet... |
| 2 | 1960 | 2 | Ray Charles | Georgia On My Mind | Georgia\|Georgia\|The whole day through\|(The whole day thr... |
| 3 | 1960 | 4 | Hank Ballard & the Midnighters | Let's Go, Let's Go, Let's Go | There's a thrill upon the hill\|Let's go, let's a-go, let's go\|Ther... |
| 4 | 1960 | 5 | Maurice Williams & the Zodiacs | Stay | (Please) Please, please, please\|Tell me you're going to\|Now y... |
| 5 | 1960 | 6 | Sam Cooke | Chain Gang | Hoh ah, hoh, I hear something saying\|Hoh ah, hoh ah\|Uh ah,... |
| 6 | 1960 | 7 | Drifters | Save The Last Dance For Me | You can dance\|Every dance with the guy\|Who gives you the ... |
| 7 | 1960 | 8 | Miracles | Shop Around | When I became of age, my mother called me to her side\|Sh... |
| 8 | 1960 | 9 | Chubby Checker | The Twist | Come on, baby, let's, do the twist\|Come on, baby, let's do th... |
| 9 | 1960 | 10 | Everly Brothers | Cathy's Clown | Don't want your love any more\|Don't want your kisses, that'... |
| 10 | 1961 | 1 | Ben E. King | Stand By Me | When the night has come\|And the land is dark\|And the moo... |
| 11 | 1961 | 2 | Patsy Cline | Crazy | Crazy, I'm crazy for feeling so lonely\|I'm crazy, crazy for feeli... |

- Augmented with data scraped from

YouTube

# Trimming

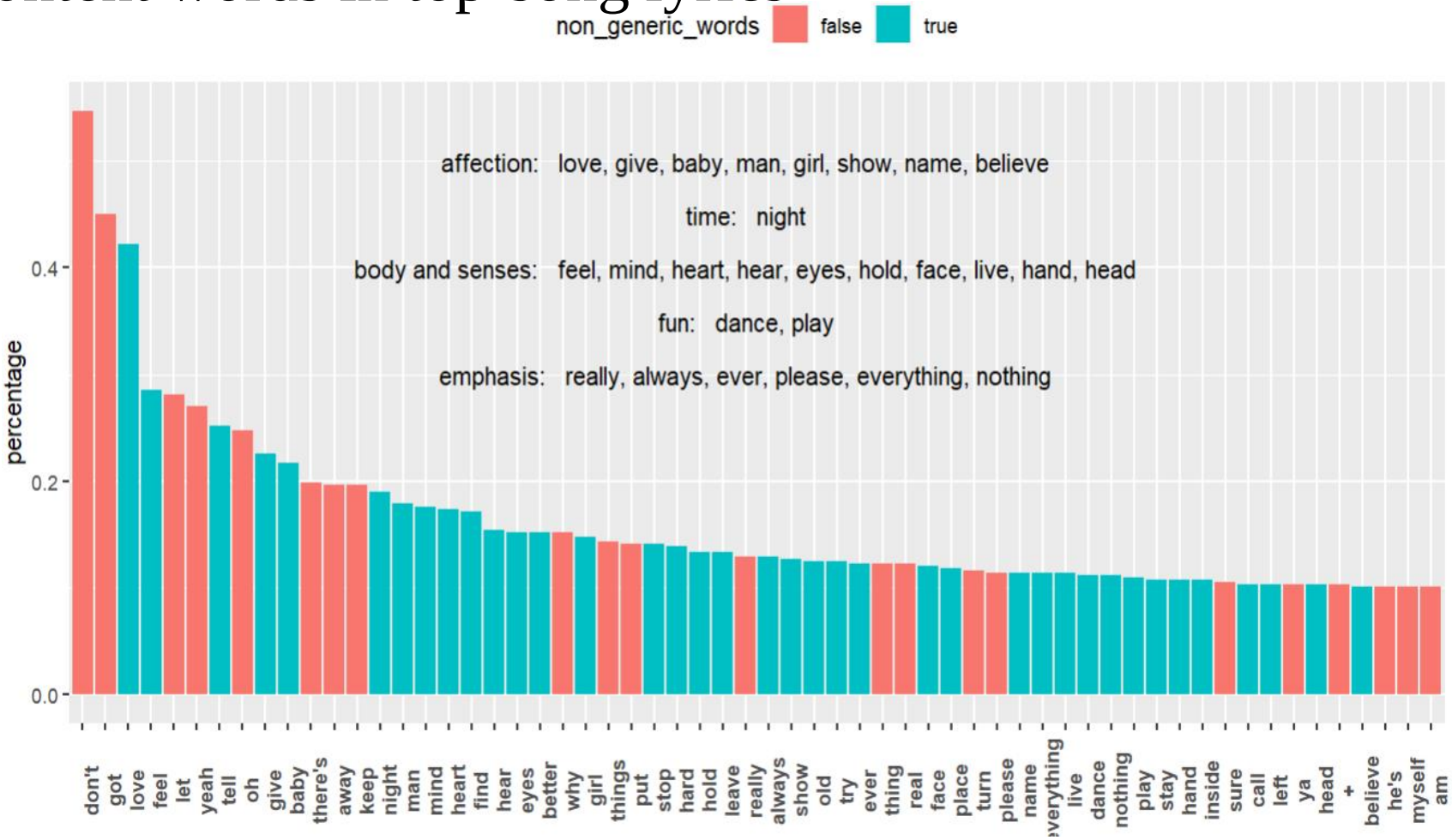
- YouTube came out in the end of 2005.
- Used only songs from 1960-2005  (remove too-old or too-new songs)
- A few songs' lyrics are incorrectly provided

| | | | | |
|---|---|---|---|---|
| 1952 | 2 | Hank Willi | Jambalaya (On The Bayou) | Goodbye Joe me gotta go me oh my oh\|Me gotta go pole the pirogue down the |
| 1952 | 3 | Dominoes | Have Mercy Baby | Have mercy, mercy baby\|I know I've done you wrong\|Have mercy, mercy baby\|I |
| 1952 | 4 | Clovers | One Mint Julep | One early morning, as I was walking\|I met a woman, we started talking\|I took he |
| 1952 | 5 | Jimmy For | Night Train | |
| 1952 | 6 | Johnny Ac | My Song | That you would leave me here in tears\|But now you're gone and hours seem like |
| 1952 | 7 | Fats Domi | Goin' Home | Can't stand your evil way\|Goin' home tomorrow\|Can't stand your evil way\|Whe |
| 1952 | 8 | King Pleas | Moody Mood For Love | There I go, there I go\|There I go, there I go\|Pretty baby, you are the soul\|Who s |
| 1952 | 9 | Little Walt | Juke | |
| 1952 | 10 | 5 Royales | Baby, Don't Do It | That you and I are through\|If you leave me pretty baby\|I'll have bread without n |
| 1953 | 1 | Drifters fe | Money Honey | Uh ooooh\|You know the landlord rang my front door bell\|I let it ring for a long, l |
| 1953 | 2 | Hank Willi | Your Cheatin' Heart | Your cheatin' heart will make you weep\|You'll cry and cry and try to sleep\|But sl |
| 1953 | 3 | Orioles | Crying In The Chapel | The tears I shed were tears of joy\|I know the meaning of contentment\|I am hap |
| 1953 | 4 | Crows | Gee | Do do-do do, do-do do, do-do do-do-do\|Do do-do do, do-do do, do-do do-do-d |
| 1953 | 5 | Faye Adam | Shake A Hand | Just leave it to me\|Don't ever be ashamed\|Just give me a chance\|I'll take care o |
| 1953 | 6 | Joe Turner | Honey Hush | In a Georgia cotton field\|Honey hush\|Come in this house, stop all that yackety y |
| 1953 | 7 | Ruth Brow | Mama, He Treats Your Daughter Mear | Mama, he treats your daughter mean\|Mama, he treats your daughter mean\|Ma |
| 1953 | 8 | Willie Mae | Hound Dog | You ain't nothin' but a hound dog\|Been snoopin' 'round my door\|You ain't nothi |
| 1953 | 9 | Hank Willi | Kaw-Liga | He fell in love with an Indian maid over in the antique store\|Kaw-Liga just stood |
| 1953 | 10 | Guitar Slim | The Things That I Used To Do | The things that I used to do\|Lord, I won't do no more\|The things that I used to d |
| 1954 | 1 | Bill Haley & | Rock Around The Clock | One, two, three o'clock, four o'clock rock\|Five, six, seven o'clock, eight o'clock r |
| 1954 | 2 | Joe Turner | Shake, Rattle And Roll | 2. Satisfaction by The Rolling Stones\|3. Imagine by John Lennon\|4. What's Going |
| 1954 | 3 | Penguins | Earth Angel | Earth angel, earth angel\|Will you be mine?\|My darling dear, love you all the time |

```
> nrow(Data)
[1] 446
```

# Lyrics analysis

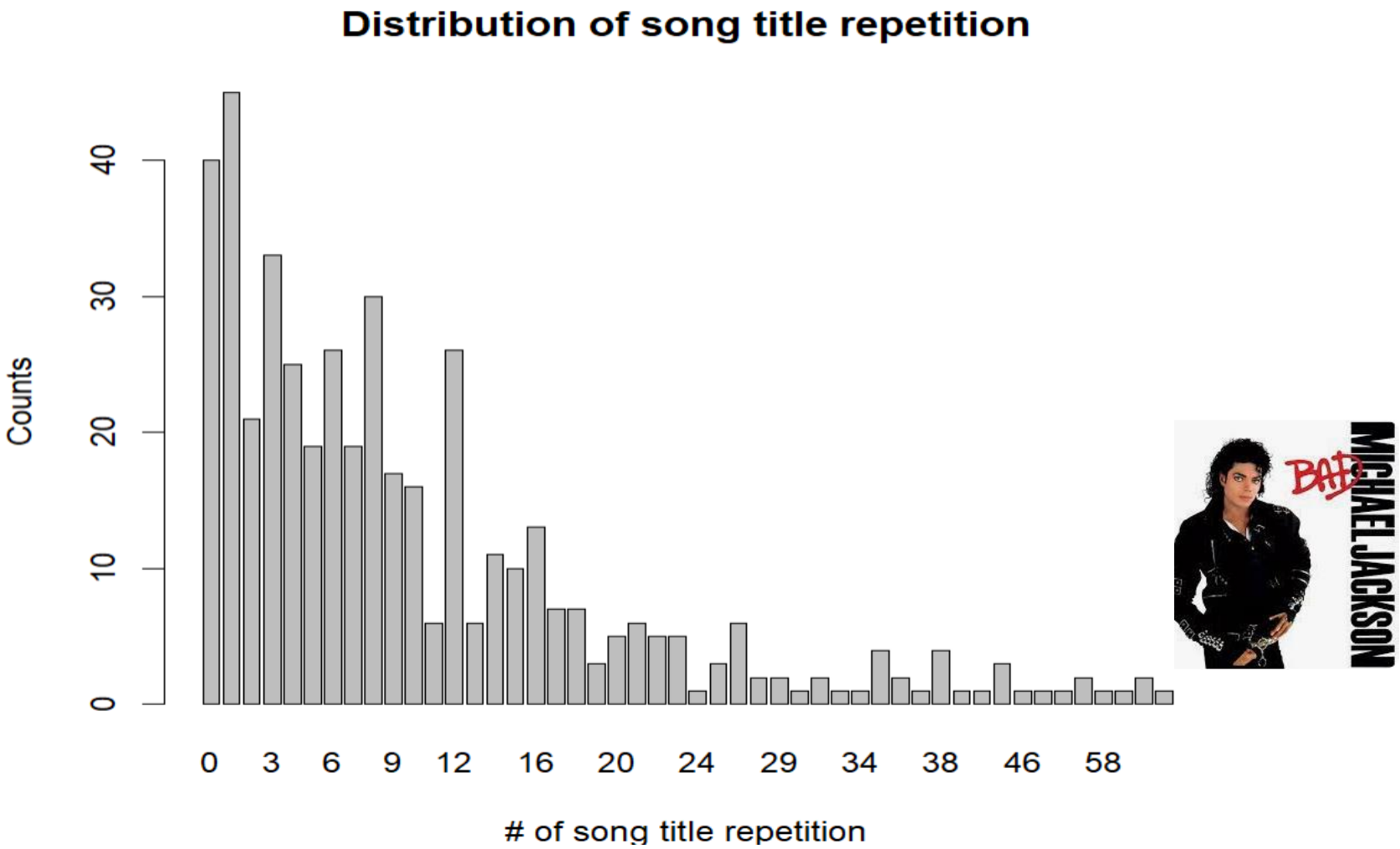
- The most common content words in top-song lyrics




- str_replace_all
- str_split
- for-loop + unique
- nested for-loop + str_detect

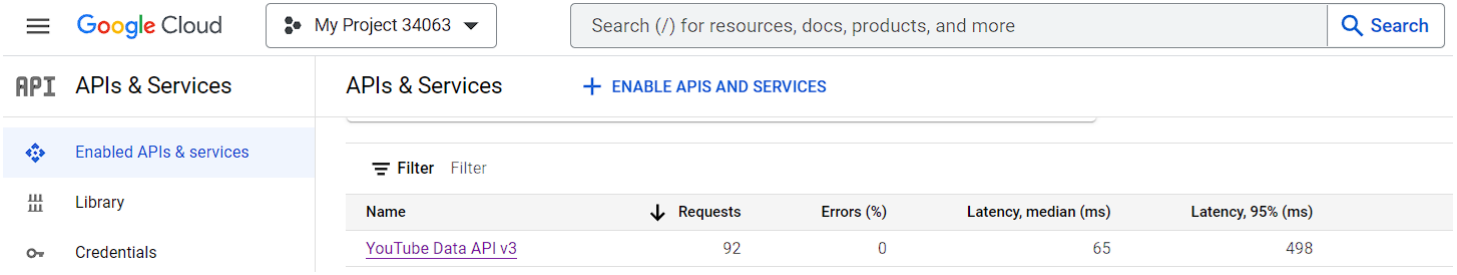
# # of times song title repeats

- str_replace_all
- tolower
- str_detect
- str_count



# Web Scraping

- **YouTube API available in Google Cloud**




youtube_api <- "https://www.googleapis.com/youtube/v3/search"

- **query string**

request_url <- paste(youtube_api , "?part=snippet&maxResults=1&q=", URLencode(search_query), "&key=", api_key, sep="")

  - part=snippet
  - maxResults=1
  - q=URLencode(search_query)
  - key="your_api_key"

| Parameters | |
|---|---|
| **Required parameters** | |
| `part` | `string`<br>The `part` parameter specifies a comma-separated list of one or more `video` resource properties that the API response will include.<br><br>If the parameter identifies a property that contains child properties, the child properties will be included in the response. For example, in a `video` resource, the `snippet` property contains the `channelId`, `title`, `description`, `tags`, and `categoryId` properties. As such, if you set `part=snippet`, the API response will contain all of those properties.<br><br>The following list contains the `part` names that you can include in the parameter value:<br>• `contentDetails`<br>• `fileDetails`<br>• `id`<br>• `liveStreamingDetails`<br>• `localizations`<br>• `player`<br>• `processingDetails`<br>• `recordingDetails`<br>• `snippet`<br>• `statistics`<br>• `status`<br>• `suggestions`<br>• `topicDetails` |

https://developers.google.com/youtube/v3/docs/videos/list

# Libraries Used in Web Scraping – httr, jsonlite

**GET() - httr**

```
# Make the API request
response <- GET(request_url)
```

**status_code() - httr**

```
# Check the HTTP status code of the response
if (status_code(response) == 200) {}
```

200 – OK; 404 – Not Found; 500 – Internal Server Error

**content() - httr**

```
# Extract content from the response
content <- content(response, "text")
```

**fromJSON() - jsonlite**

```
# Parse JSON content
video_details <- fromJSON(content)
```

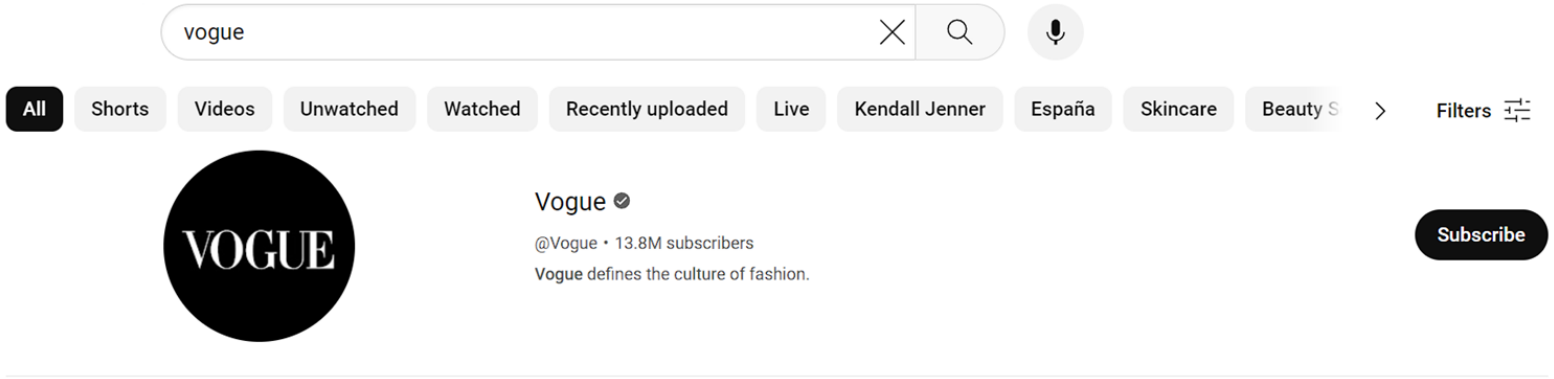
| | | |
|---|---|---|
| video_details | list [6] | List of length 6 |
| kind | character [1] | 'youtube#videoListResponse' |
| etag | character [1] | 'ALmgp32eScFEfjP58HzLuV_VqmU' |
| items | list [1 x 4] (S3: data.frame) | A data.frame with 1 row and 4 columns |
| kind | character [1] | 'youtube#video' |
| etag | character [1] | 'Pb3rK8ZpjyUQG0X2OOAmdBQjF1M' |
| id | character [1] | 'Q1dUDzBdnmI' |
| statistics | list [1 x 4] (S3: data.frame) | A data.frame with 1 row and 4 columns |
| viewCount | character [1] | '140220452' |
| likeCount | character [1] | '721342' |
| favoriteCount | character [1] | '0' |
| commentCount | character [1] | '23462' |
| pageInfo | list [2] | List of length 2 |
| totalResults | integer [1] | 1 |
| resultsPerPage | integer [1] | 1 |
| search_query | character [1] | 'Check On It Beyonc?? feat. Slim Thug' |
| video_id | character [1] | 'Q1dUDzBdnmI' |

**write_json() - jsonlite**

```
# Write json file
write_json(json, “data.json", pretty = TRUE, auto_unbox = TRUE)
```

# Web Scraping Tips

- Sometimes, the first search result is not a video but an advertisement.




- Special symbols do not work and can cause errors in search queries, for example, when searching for a song named 'Neighborhood #1 (Tunnels)'.

- The problem was then solved by using a combined query with both the **song name** and the **singer's name**. This approach makes sense, as some song names can refer to other things.

# Regular Expression in Handling Json

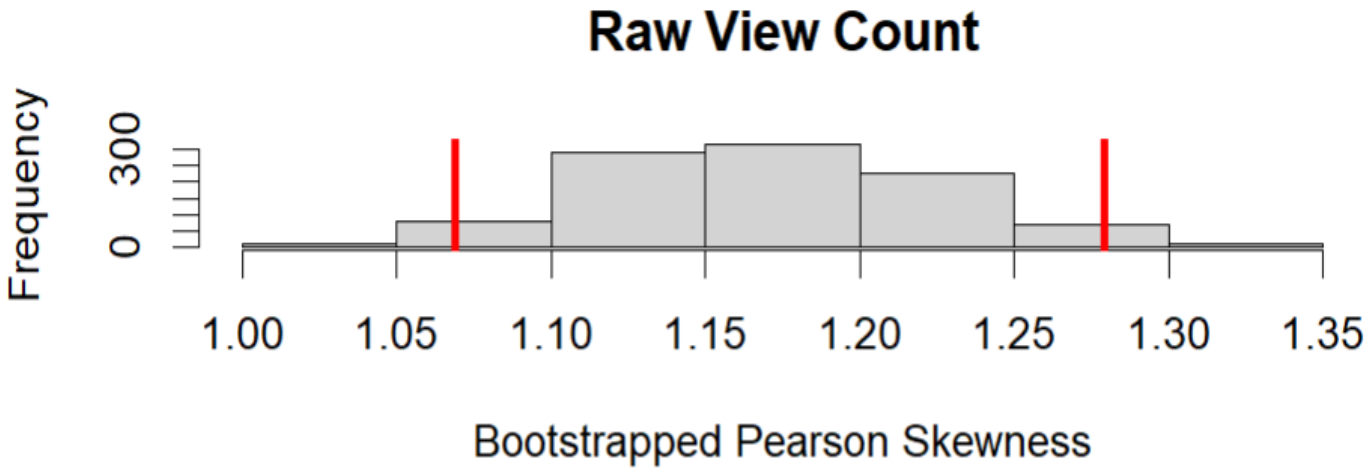
- We created the regular expression pattern for the features we want, including (song_name, view_count, like_ count)

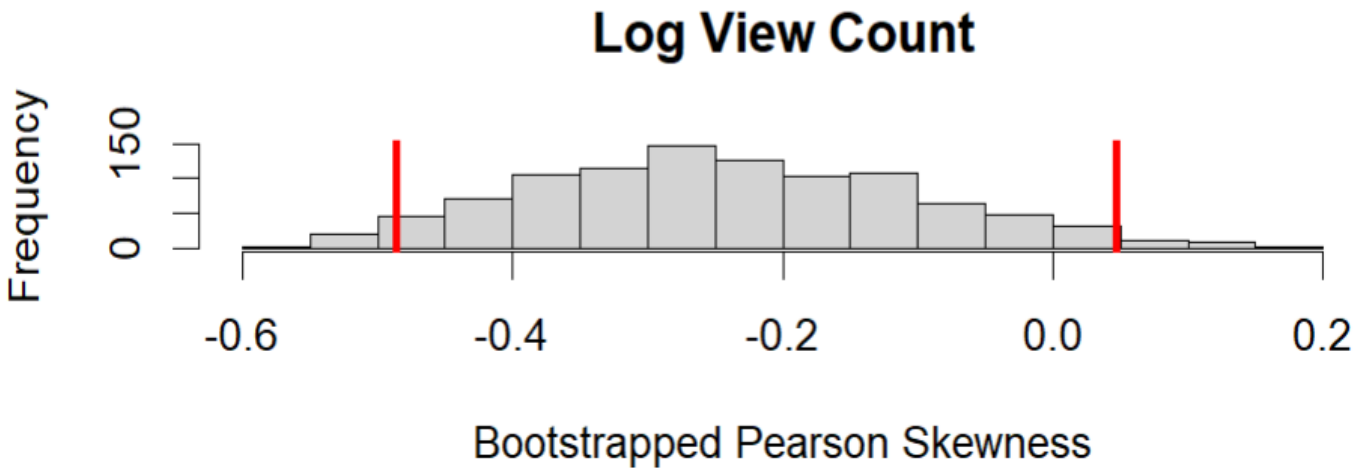
- The function we used is str_extract_all() to find all matching pairs.

| search_query | viewCount | likeCount |
|---|---|---|
| Will You Love Me Tomorrow Shirelles | 11038942 | 92542 |
| Georgia On My Mind Ray Charles | 2077198 | 28663 |
| Let's Go, Let's Go, Let's Go Hank Ballard & the Midnig... | 647485 | 4972 |
| Stay Maurice Williams & the Zodiacs | 9269625 | 56605 |
| Chain Gang Sam Cooke | 8136356 | 72602 |
| Save The Last Dance For Me Drifters | 35736750 | 160597 |
| Shop Around Miracles | 359871 | 3445 |
| The Twist Chubby Checker | 5935430 | 39116 |
| Cathy's Clown Everly Brothers | 3833138 | 26571 |
| Stand By Me Ben E. King | 9899281 | 91611 |
| Crazy Patsy Cline | 19023476 | 204739 |
| The Wanderer Dion | 32245658 | 298597 |
| Runaround Sue Dion | 30484544 | 297316 |
| Crying Roy Orbison | 7695729 | 56769 |
| Hit The Road Jack Ray Charles | 17898477 | 319281 |
| Runaway Del Shannon | 7875941 | 81034 |
| Quarter To Three Gary U.S. Bonds | 264328 | 2592 |
| It Will Stand Showmen | 506756 | 2134 |
| Running Scared Roy Orbison | 677486 | 7518 |
| Bring It On Home To Me Sam Cooke | 5419440 | 60149 |
| You've Really Got A Hold On Me Miracles | 10804387 | 105068 |
| The Loco-Motion Little Eva | 23747401 | 134406 |
| Sherry Four Seasons | 6716155 | 65981 |

# Skewness of the Data Distribution Using Bootstrap

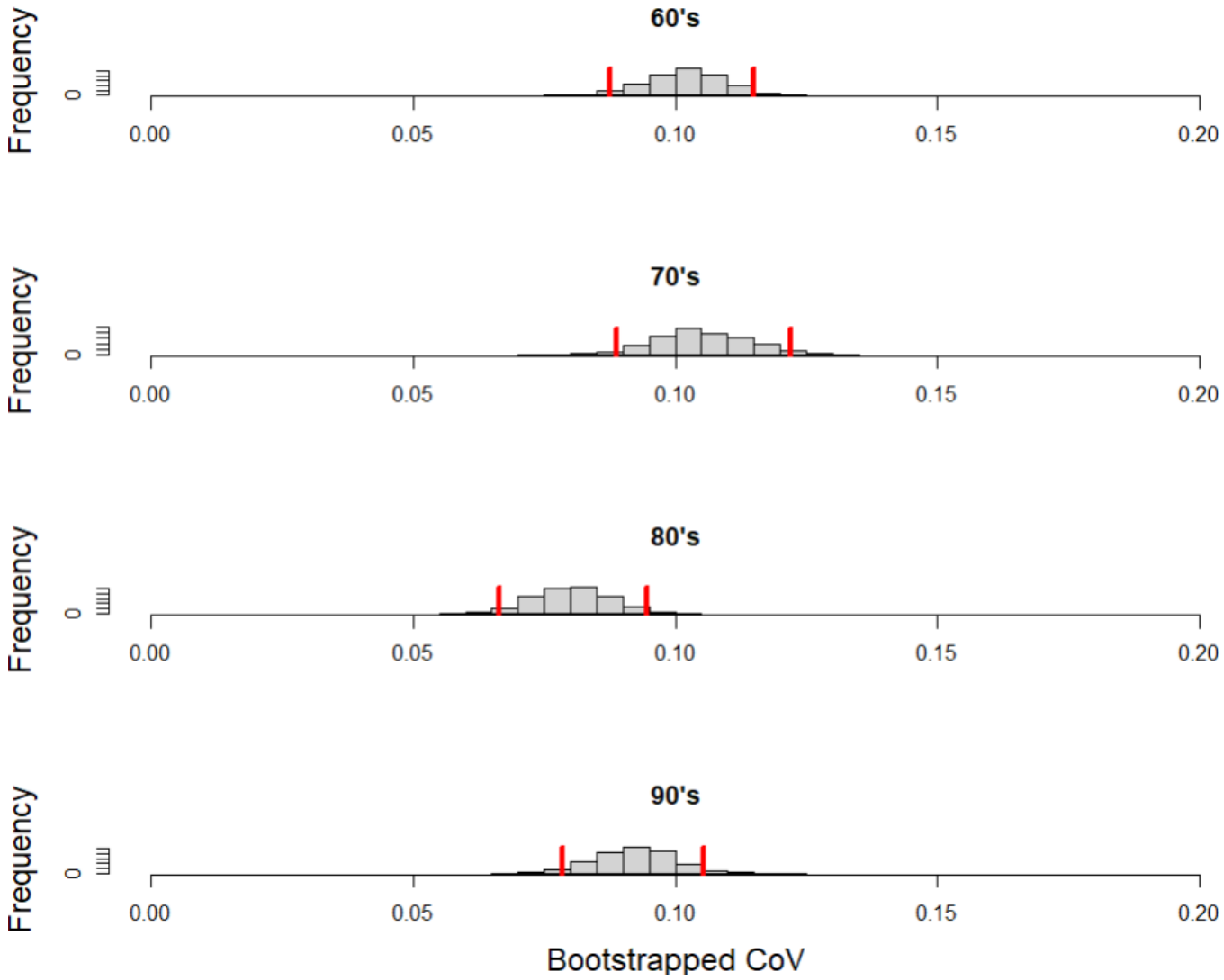
- Perform a bootstrap analysis to estimate the distribution of the Pearson's median skewness coefficient for the view_count variable.
- This histogram has values around **1.05 to 1.3**, indicating a **positive skew** in the view_count data.
- The skewness values after the log transformation range from about **-0.5 to 0.05**. The distribution is more centered around zero, suggesting that log transformation can help reduce the skewness of the data, leading to a more **symmetric** distribution.



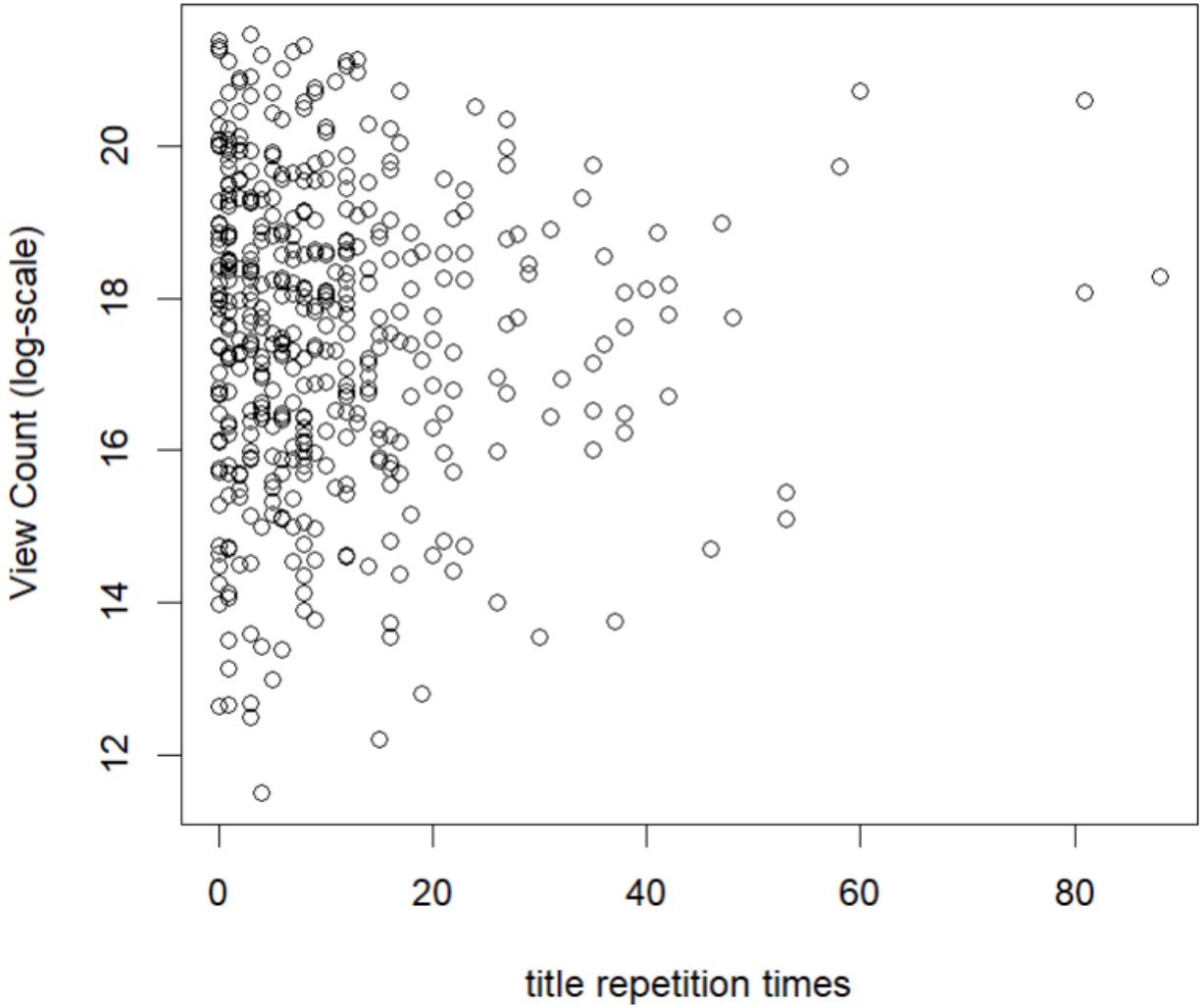
# Analysis on Coefficient of Variation

$$\text{Coefficient of Variation (CV)} = \frac{\text{Standard Deviation (s)}}{\text{Sample Mean } (\bar{x})}$$




- We used bootstrap to get an interval **for each decade**.
- The CV for 80's is comparatively small, which indicates greater consistency and less variability in the view counts of songs from the 80’s.

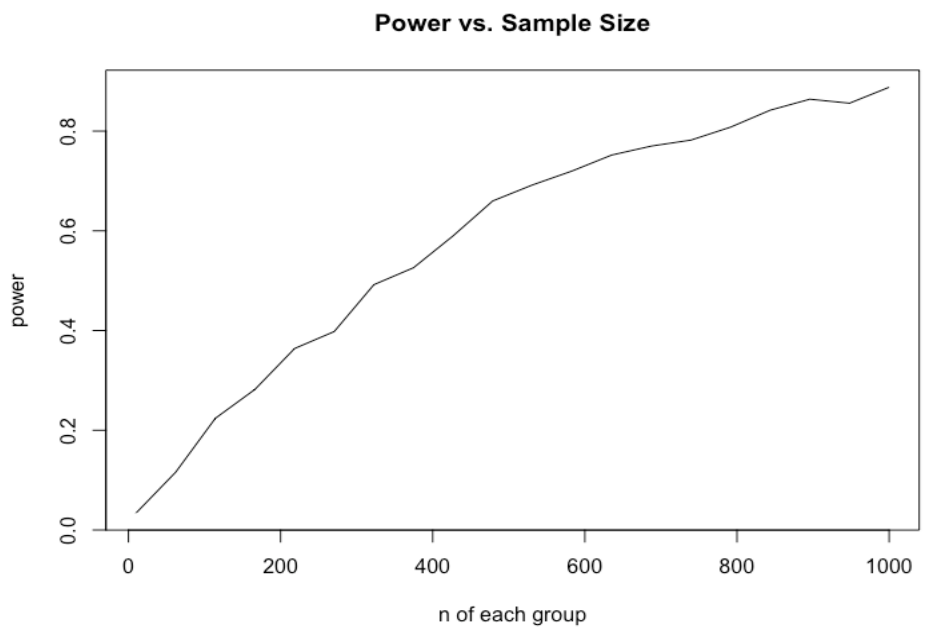
# View Count ~ title repetition times
# Choice of predictor (continuous or categorical?)




| | title_times_category | sample mean (log-ViewCount) | sample sd | sample size |
|---|---|---|---|---|
| 1 | 0 time | 17.62835 | 2.171227 | 40 |
| 2 | 1 time | 17.59714 | 2.142629 | 45 |
| 3 | 2 times | 18.02457 | 2.065578 | 21 |
| 4 | 3--4 times | 17.36523 | 2.105227 | 58 |
| 5 | 5--8 times | 17.42042 | 1.883426 | 94 |
| 6 | more than 8 times | 17.58930 | 1.842117 | 188 |

# Monte Carlo Power Study

- We applied Monte Carlo method for the Power Study
- Assumptions:
  - In each group (by repetition times), the sample size n is the same
  - The true distribution has mean and standard deviation same as our sample mean and standard deviation
- F test: aov()
- Power becomes high when sample size is at least 800



# Parametric F test(AOV)

Null hypothesis($H_0$): The means of each group are the same

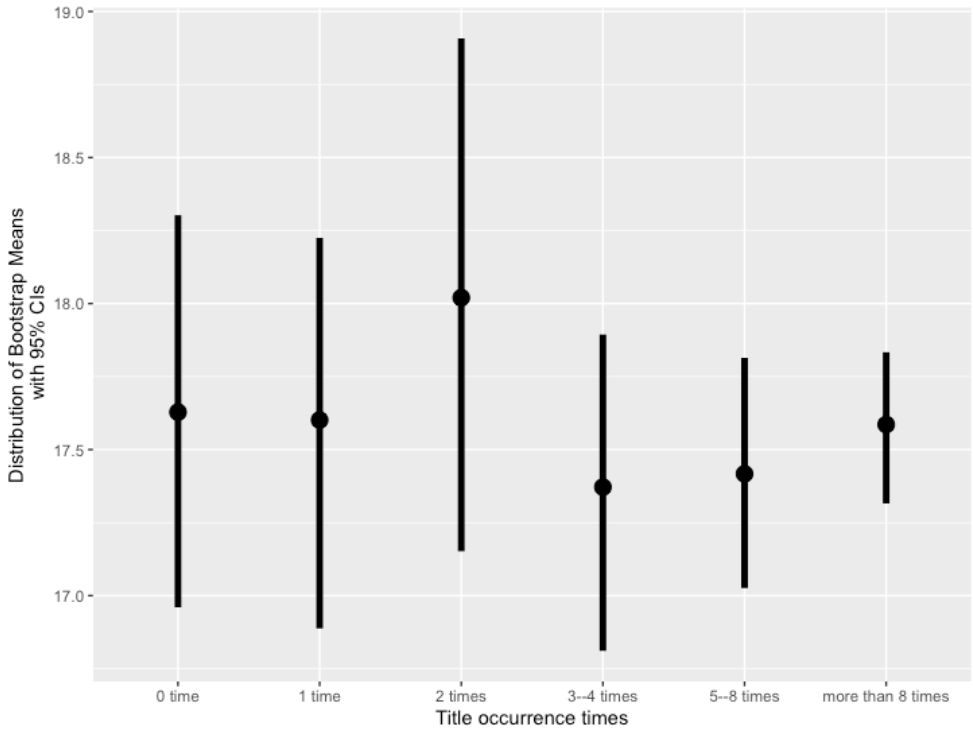
Alternative hypothesis($H_\alpha$): At least one sample mean differs from the others.

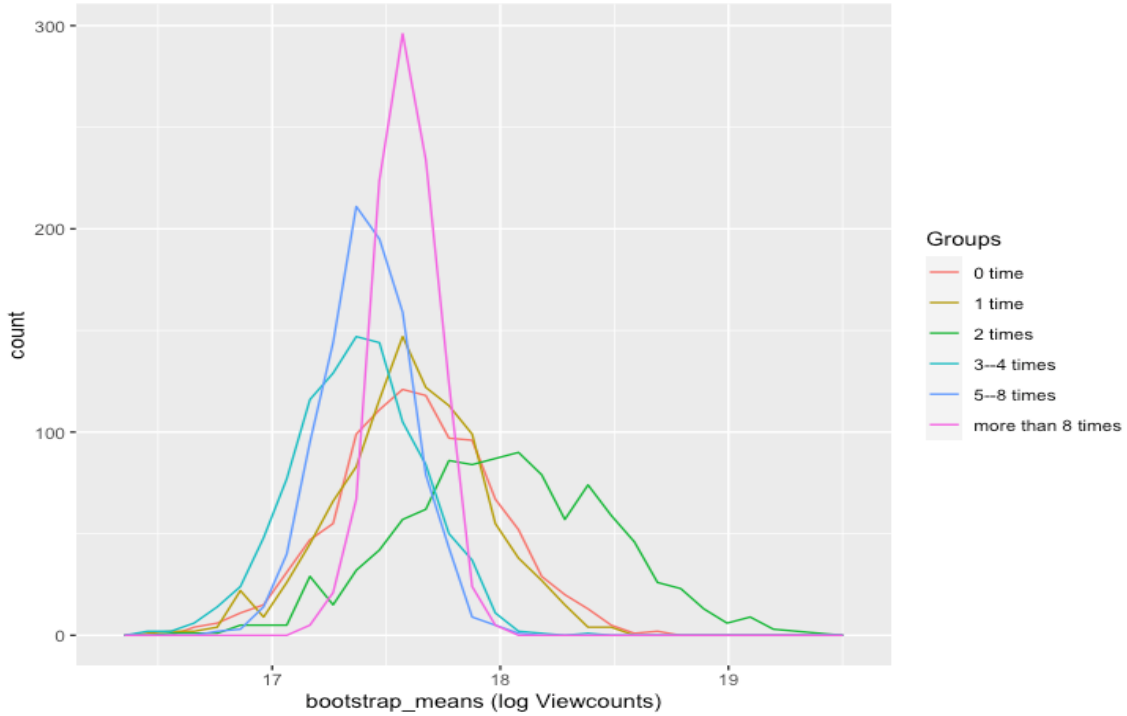
Result: 0.802>0.05, fail to reject the null hypothesis

```
> summary(aov(log(Data$View_Count)~Data$title_times_categorical))
                                Df Sum Sq Mean Sq F value Pr(>F)
Data$title_times_categorical     5    8.9   1.785   0.465  0.802
Residuals                      440 1688.3   3.837
```

# Further Validation

- Bootstrap for creating confidence interval for mean of each group
- From the graph, it shows that there is no big differences among all groups, which aligns with our F test and Power study (we might be **underpowered**).



# Conclusion

- There is no linear relationship between the occurrence of song title in lyrics and the popularity of songs

- Our analysis shows with a slight evidence that maybe having occurrence of 2 times might be the better option in song.