Collecting and Analyzing Billboard Chart Data from Wikipedia

Róbert Kendereš, DAV2

June 11, 2024

1 Introduction

This report presents an analysis of Billboard chart data taken from Wikipedia, focusing on the top artists, genres, and song characteristics since the creation of Billboard charts.

2 Dataset

Assembling the dataset was one of the more demanding tasks of this project, since the data was scattered over a group of multiple pages of Wikipedia, nested in one another.
We worked with web crawlers, specifically using the python library **scrapy** to get all the information we need from the html code of the sites.

On the first site we can see a list of years where the Billboard charts were present, each year embedded with a link to another site, which contains the actual chart of top 10 most popular songs of that year. Each of these songs has a link to a Wikipedia page of that given song, from which we can get more information about it, namely date of release, genre and song duration.

3 Acquiring Data

To begin, we utilized the `SpiderWiki` spider to crawl through Wikipedia pages, starting with the "Lists of Billboard number-one singles" page. This page listed the years when the Billboard charts were active, each linked to another Wikipedia page with the top 10 charting songs for that year.
Using the `parse` method in `SpiderWiki`, we extracted information such as the year, era (pre-Hot100 or not), and URL of each Billboard chart list.

After storing the results of this spiderWiki crawl into **billboard1.csv**, we then followed these year-URLs to individual yearly pages using the second spider, spiderBillboard, to gather data about individual songs, including titles, artists, and links to their respective Wikipedia pages.
In the creation process of the tables we noticed some inconsistencies between the tables on the wikipedia pages for different years, hence we needed to add some more filtering, as with pre and post 2012 top 10 lists:

```
def parse(self, response):
    if response.meta['pre-2012'] == 'True':
        table = response.xpath("(//table[@class='wikitable'])[2]//tr")
    else:
        table = response.xpath("(//table[@class='wikitable plainrowheaders'])[1]//tr")
```

The results of the second crawl were then stored in **billboard2.csv.**
After this, cleanup.py creates a table which is more suitable for the project, by adding a column of number of occurences of the given song on the Billboard charts since 1959, and removing other occurences of this song in the table.
Following this we run the final spider, **spiderSong**. This spider visits a specific table on each song's site which contains the information we need to construct our final dataset, and stores it in **songs.csv**.

Through this process, we collected essential data attributes from Wikipedia pages, which formed the basis for our subsequent analysis.

4 Analysis

4.1 Top Genres by Occurrences on Billboard

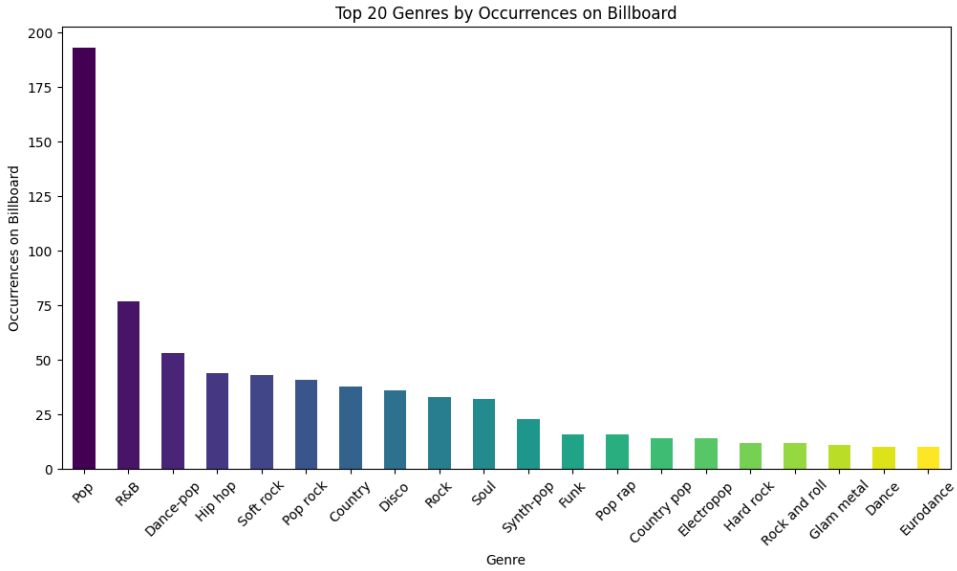

Figure 1: Top 20 Genres by Number of Occurrences on Billboard

The histogram in Figure 5 illustrates the top 20 genres by occurrences on the Billboard charts. As expected, the Pop (Popular music) genre dominates the charts, surpassing other genres by a large amount, and

4.2 Median Duration of Songs on Billboard

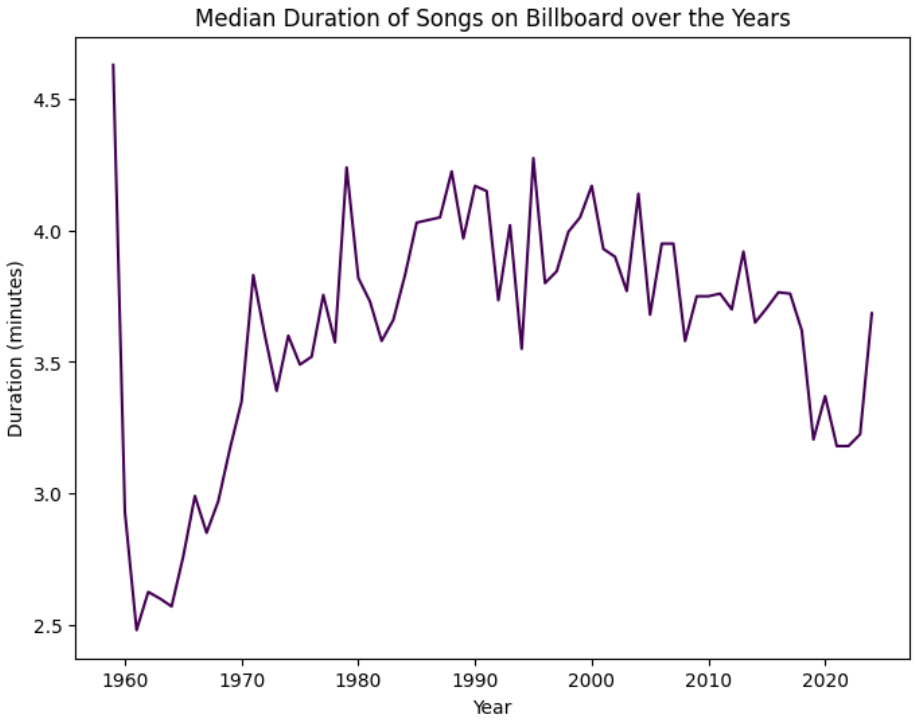

Figure 2: Median Duration of Songs on Billboard over the Years

Figure 5 depicts the median duration of songs on the Billboard charts over the years. We observe a stabilization around the four-minute mark around 1990, followed by a decreasing trend in song duration in the early 21st century.

4.3 Top Artists by Occurrences on Billboard

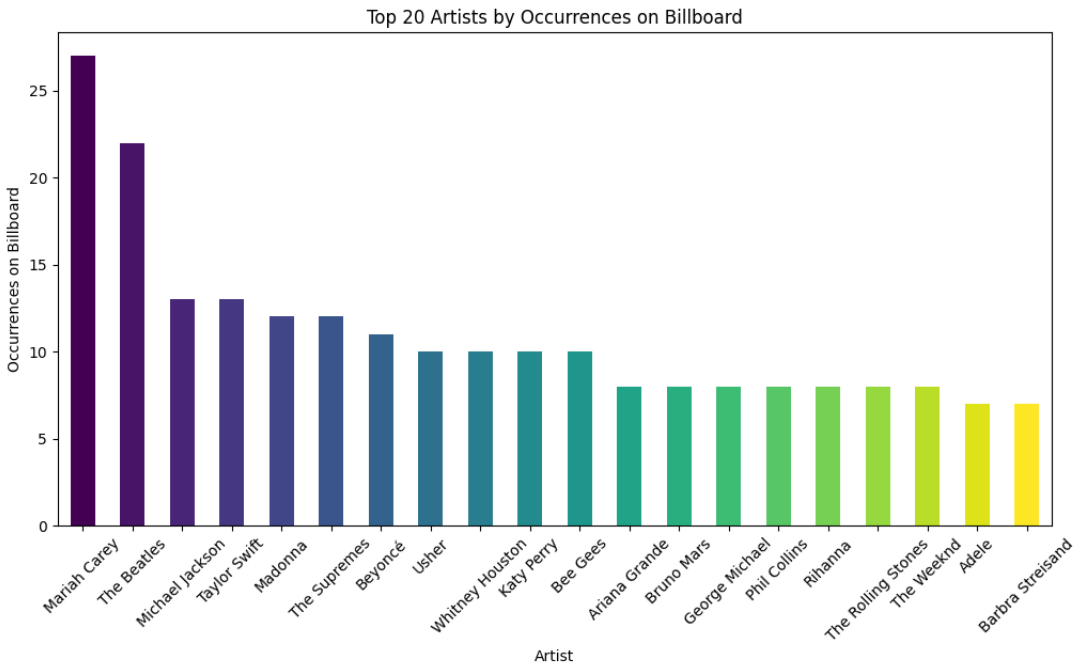

Figure 3: Top 20 Artists by Occurrences on Billboard

The chart in Figure 5 showcases the top 20 artists by occurrences on the Billboard charts. Mariah Carey and The Beatles emerge as the leading artists with 27 and 22 occurrences, respectively, followed by Michael Jackson with 13 occurrences.

4.4 Duration of Songs per Genre

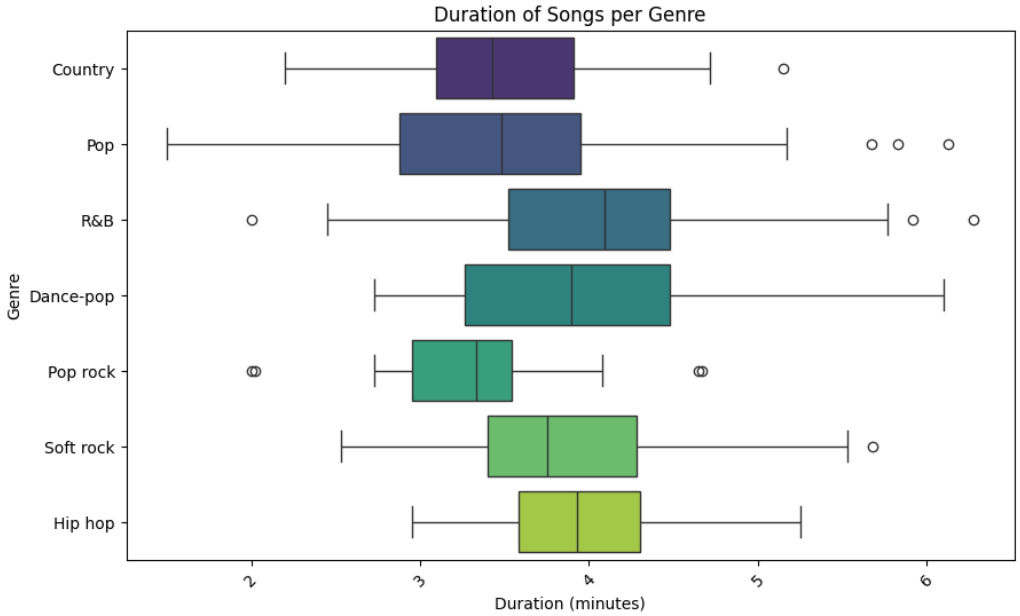

Figure 4: Duration of Songs per Genre

Figure 5 presents the duration distribution of songs across different genres on the Billboard charts. The majority of songs across genres have durations approximately around and under four minutes.

4.5 Most Popular Genre per Year

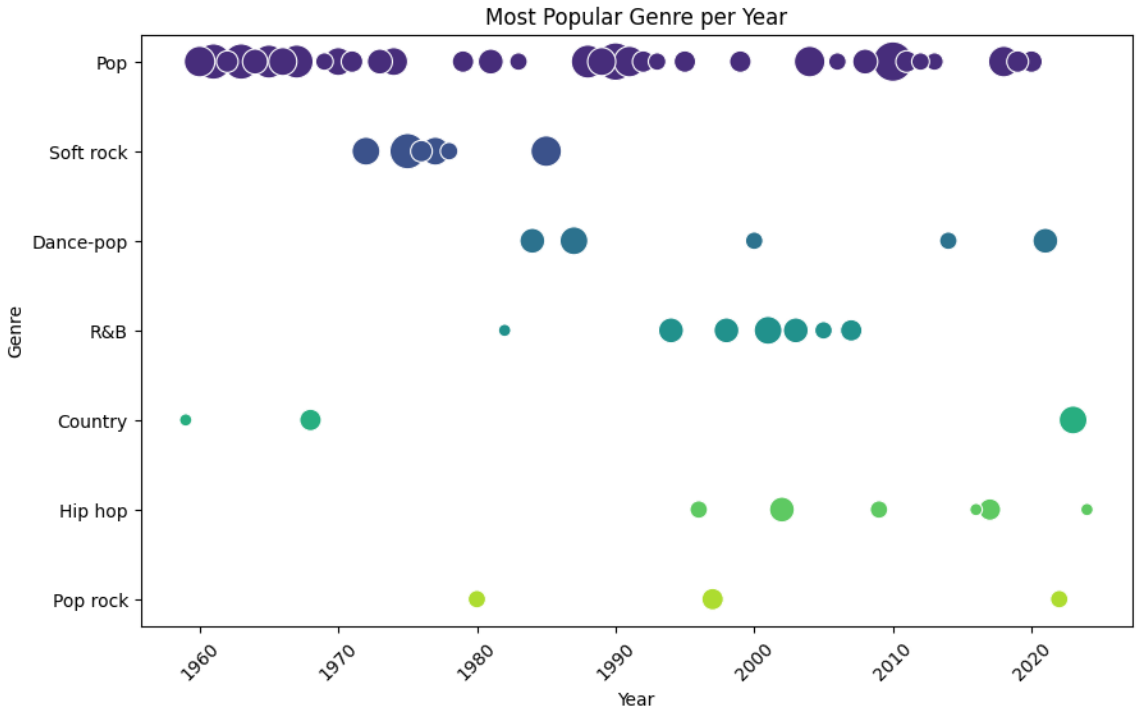

Figure 5: Most Popular Genre per Year

Finally, Figure 5 displays the most popular genre per year on the Billboard charts. Over the years, we observe fluctuations in genre popularity, but Pop seems to always come back. Starting with Pop, before 1980 cam a Soft rock era, then Dance-pop and Disco era, R&B was the main music of the 2000s, occasionally switching with Hip-Hop, and the last couple of years we can see more of each genre. From the chart above we could conclude that in the last couple of years, people are becoming more open with their music tastes.

References

https://docs.scrapy.org/en/latest/topics/spiders.html
Lists of Billboard number one singles
List of Billboard Hot 100 number ones of 1959
python virtual environment