Temporal Trends in Popular Music Genres: A Data-Driven Analysis

Elizabeth Hall - Data Analysis Project

Elizabeth Hall

2024-02-09

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

# 1. Part 1

## 1.1 The Data

The data I have so far is a list of the Billboard Hot 100s from 1958 to 2023. (Which can be found here <https://github.com/HipsterVizNinja/random-data/blob/main/Music/hot-100/README.md>)

The data can also be found in the “raw-data” folder in this GitHub repo. It is titled “Hot\_100s.csv”.

This dataset includes observations for chart position, chart date, song name, performer name, song id, instance, time on chart, consecutive weeks, previous week, peak position, worst position, chart debut date, and chart url. The data was collected from the Billboards Hot 100

I will be adding the genre category to this dataset myself by cross-referencing the artists with the Spotify API (found here -<https://developer.spotify.com/documentation/web-api>) to obtain their musical genre. Though I am not entirely sure how the data was collected, I know that the Billboard Hot 100s keeps weekly records on their website, so this is where the data came from.

Here is some code to display a little info on the data:

chart\_position chart\_date song performer  
1 84 1990-05-05 "B" Girls Young And Restless  
2 78 1990-05-12 "B" Girls Young And Restless  
3 68 1990-05-19 "B" Girls Young And Restless  
4 60 1990-05-26 "B" Girls Young And Restless  
5 58 1990-06-02 "B" Girls Young And Restless  
6 55 1990-06-09 "B" Girls Young And Restless  
 song\_id instance time\_on\_chart consecutive\_weeks  
1 "B" GirlsYoung And Restless 1 1 NA  
2 "B" GirlsYoung And Restless 1 2 1  
3 "B" GirlsYoung And Restless 1 3 2  
4 "B" GirlsYoung And Restless 1 4 3  
5 "B" GirlsYoung And Restless 1 5 4  
6 "B" GirlsYoung And Restless 1 6 5  
 previous\_week peak\_position worst\_position chart\_debut  
1 NA 84 84 1990-05-05  
2 84 78 84 1990-05-05  
3 78 68 84 1990-05-05  
4 68 60 84 1990-05-05  
5 60 58 84 1990-05-05  
6 58 55 84 1990-05-05  
 chart\_url  
1 https://www.billboard.com/charts/hot-100/1990-05-05  
2 https://www.billboard.com/charts/hot-100/1990-05-12  
3 https://www.billboard.com/charts/hot-100/1990-05-19  
4 https://www.billboard.com/charts/hot-100/1990-05-26  
5 https://www.billboard.com/charts/hot-100/1990-06-02  
6 https://www.billboard.com/charts/hot-100/1990-06-09

Data Summary:

'data.frame': 338395 obs. of 13 variables:  
 $ chart\_position : int 84 78 68 60 58 55 54 57 67 75 ...  
 $ chart\_date : chr "1990-05-05" "1990-05-12" "1990-05-19" "1990-05-26" ...  
 $ song : chr "\"B\" Girls" "\"B\" Girls" "\"B\" Girls" "\"B\" Girls" ...  
 $ performer : chr "Young And Restless" "Young And Restless" "Young And Restless" "Young And Restless" ...  
 $ song\_id : chr "\"B\" GirlsYoung And Restless" "\"B\" GirlsYoung And Restless" "\"B\" GirlsYoung And Restless" "\"B\" GirlsYoung And Restless" ...  
 $ instance : num 1 1 1 1 1 1 1 1 1 1 ...  
 $ time\_on\_chart : int 1 2 3 4 5 6 7 8 9 10 ...  
 $ consecutive\_weeks: num NA 1 2 3 4 5 6 7 8 9 ...  
 $ previous\_week : num NA 84 78 68 60 58 55 54 57 67 ...  
 $ peak\_position : int 84 78 68 60 58 55 54 54 54 54 ...  
 $ worst\_position : int 84 84 84 84 84 84 84 84 84 84 ...  
 $ chart\_debut : chr "1990-05-05" "1990-05-05" "1990-05-05" "1990-05-05" ...  
 $ chart\_url : chr "https://www.billboard.com/charts/hot-100/1990-05-05" "https://www.billboard.com/charts/hot-100/1990-05-12" "https://www.billboard.com/charts/hot-100/1990-05-19" "https://www.billboard.com/charts/hot-100/1990-05-26" ...

Descriptive Statistics:

chart\_position chart\_date song performer   
 Min. : 1.0 Length:338395 Length:338395 Length:338395   
 1st Qu.: 25.5 Class :character Class :character Class :character   
 Median : 50.0 Mode :character Mode :character Mode :character   
 Mean : 50.5   
 3rd Qu.: 75.0   
 Max. :100.0   
   
 song\_id instance time\_on\_chart consecutive\_weeks  
 Length:338395 Min. : 1.000 Min. : 1.000 Min. : 1.00   
 Class :character 1st Qu.: 1.000 1st Qu.: 4.000 1st Qu.: 3.00   
 Mode :character Median : 1.000 Median : 7.000 Median : 7.00   
 Mean : 1.086 Mean : 9.251 Mean : 8.79   
 3rd Qu.: 1.000 3rd Qu.:13.000 3rd Qu.:12.00   
 Max. :15.000 Max. :91.000 Max. :89.00   
 NA's :33667   
 previous\_week peak\_position worst\_position chart\_debut   
 Min. : 1.00 Min. : 1.00 Min. : 1.00 Length:338395   
 1st Qu.: 23.00 1st Qu.: 13.00 1st Qu.: 74.00 Class :character   
 Median : 47.00 Median : 38.00 Median : 86.00 Mode :character   
 Mean : 47.56 Mean : 40.77 Mean : 80.64   
 3rd Qu.: 71.00 3rd Qu.: 65.00 3rd Qu.: 94.00   
 Max. :100.00 Max. :100.00 Max. :100.00   
 NA's :33667   
 chart\_url   
 Length:338395   
 Class :character   
 Mode :character

First few rows of the Data Frame:

chart\_position chart\_date song performer  
1 84 1990-05-05 "B" Girls Young And Restless  
2 78 1990-05-12 "B" Girls Young And Restless  
3 68 1990-05-19 "B" Girls Young And Restless  
4 60 1990-05-26 "B" Girls Young And Restless  
5 58 1990-06-02 "B" Girls Young And Restless  
6 55 1990-06-09 "B" Girls Young And Restless  
 song\_id instance time\_on\_chart consecutive\_weeks  
1 "B" GirlsYoung And Restless 1 1 NA  
2 "B" GirlsYoung And Restless 1 2 1  
3 "B" GirlsYoung And Restless 1 3 2  
4 "B" GirlsYoung And Restless 1 4 3  
5 "B" GirlsYoung And Restless 1 5 4  
6 "B" GirlsYoung And Restless 1 6 5  
 previous\_week peak\_position worst\_position chart\_debut  
1 NA 84 84 1990-05-05  
2 84 78 84 1990-05-05  
3 78 68 84 1990-05-05  
4 68 60 84 1990-05-05  
5 60 58 84 1990-05-05  
6 58 55 84 1990-05-05  
 chart\_url  
1 https://www.billboard.com/charts/hot-100/1990-05-05  
2 https://www.billboard.com/charts/hot-100/1990-05-12  
3 https://www.billboard.com/charts/hot-100/1990-05-19  
4 https://www.billboard.com/charts/hot-100/1990-05-26  
5 https://www.billboard.com/charts/hot-100/1990-06-02  
6 https://www.billboard.com/charts/hot-100/1990-06-09

I may also use other datasets to supplement or do further analysis but I do not have those currently.

## 1.2 The Question

The question I want to answer is: How has the popularity of different music genres evolved over time?

The primary outcome of interest is to see what if any temporal trends exist in the data. Ex: Was pop less popular in the 50-90s and more popular in the 2000s? Did rock become less popular after the 80s?

Some specific predictors I want to look at are time(years), genre, and artists. I may also look into societal and cultural events, or outside factors that could influence music trends. This would not be 100% reliable, but it would be interesting to see if any events line up with trends in the data. The main relationships/patterns that I will be looking for in the data are patterns/trends in genre popularity over time. Though if I find more interesting datasets this may expand and change some.

## 1.3 Analysis

I will likely be analyzing time(years) vs genre vs popularity. Popularity will be some measure based on time spent on chart, peak position, and consecutive weeks. I am not 100% sure how this system will work yet but that is the vague idea. Genre analysis, as mentioned earlier, will be done using the Spotify API (found here - <https://developer.spotify.com/documentation/web-api>) to obtain genre per artist.

# 2. Summary/Abstract

*Write a summary of your project.*

# 3. Introduction

## 3.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 3.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

## 3.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above (here dataanalysis\_template\_references.bib) and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020).

# 4. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 4.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). **?@fig-schematic** is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

## 4.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 4.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 4.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 5. Results

## 5.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

**?@tbl-summarytable** shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below.

## 5.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

**?@fig-result** shows a scatterplot figure produced by one of the R scripts.

## 5.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example **?@tbl-resulttable2** shows a summary of a linear model fit.

# 6. Discussion

## 6.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 6.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 6.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, et al., 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 7. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>