Top Music Genres

Jenna Moscaritolo

December 10, 2020

Abstract

In this project, we will look at top 500 albums from the years 1955-2011 which was used from The Rolling Stone magazine. We will be looking at the probabilities of each genre based on which genre is the most popular in each year. Below is a bar graph of the different genres where, after data cleaning, there is only one genre per year (2009 has no data). For the initial data, look at Figure 1 in the Appendix.

Introduction

The Rolling Stone is a monthly magazine that comments on ideas of popular culture. Founded in San Francisco, California, it was initially known for its coverage of rock music, but in the 1990s, the company expanded its strict range of music to popular entertainment forms such as television shows and actors. Nowadays, it even covers politics.

The dataset from The Rolling Stone has six column vectors: Number (number album is listed), Year (year album was released), Album (album name), Artist (artist name), Genre (main genre of album), and Subgenre (extra genre information of album). We will only be looking at Year and Genre for this project.

The Year values range from 1955 to 2011 with no data for 2009 as there are no entries, meaning that there must not have been a top 500 album in 2009. After looking at the raw Genre data, I decided to use just the first word of each entry. The 13 Genre categories are Blues, Jazz, Classical, Reggae, World, Pop, Stage, Folk, Latin, Electronic, Funk, Hip, and Rock (Note: World includes country, Stage includes screen, and Folk includes soul).

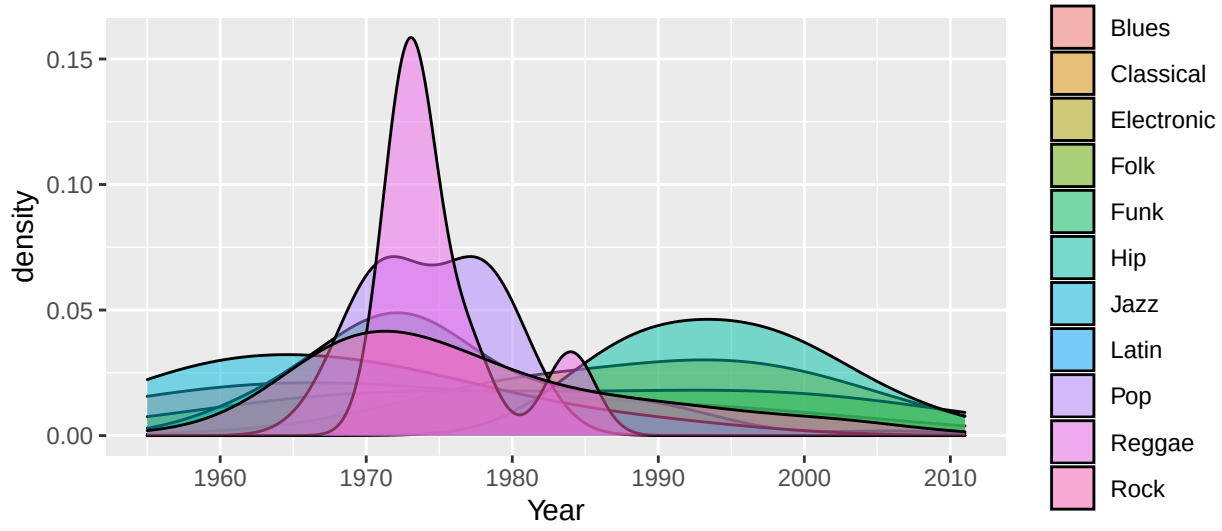

Figure 1: Initial EDA: Density Plot

This density plot shows how often each genre is recorded for each year of the dataset (1955-2011). Although it is sometimes tough to distinguish the colors, it is clear that Rock was the main genre, by far, around the years 1971 to 1976 and around a similar group of years (about 1968-1982), pop became popular. It seems that around 1985 to 2008, the most popular genre was Hip Hop.

The main questions that I will be answering in this report are what will the top three categories be over the next few years and how will the popularity of the different categories change over time.

Method

I started with a multinomial regression because I knew the data is categorical. So, I fit this model using multinom() and vglm(). I decided to stick with multinom() because vglm() had more information that was not relevant to the project. The function multinom() includes the coefficients and the standard errors of each Genre (other than Blues as that is considered the baseline).

Here is a scatterplot of the Genres over the years.

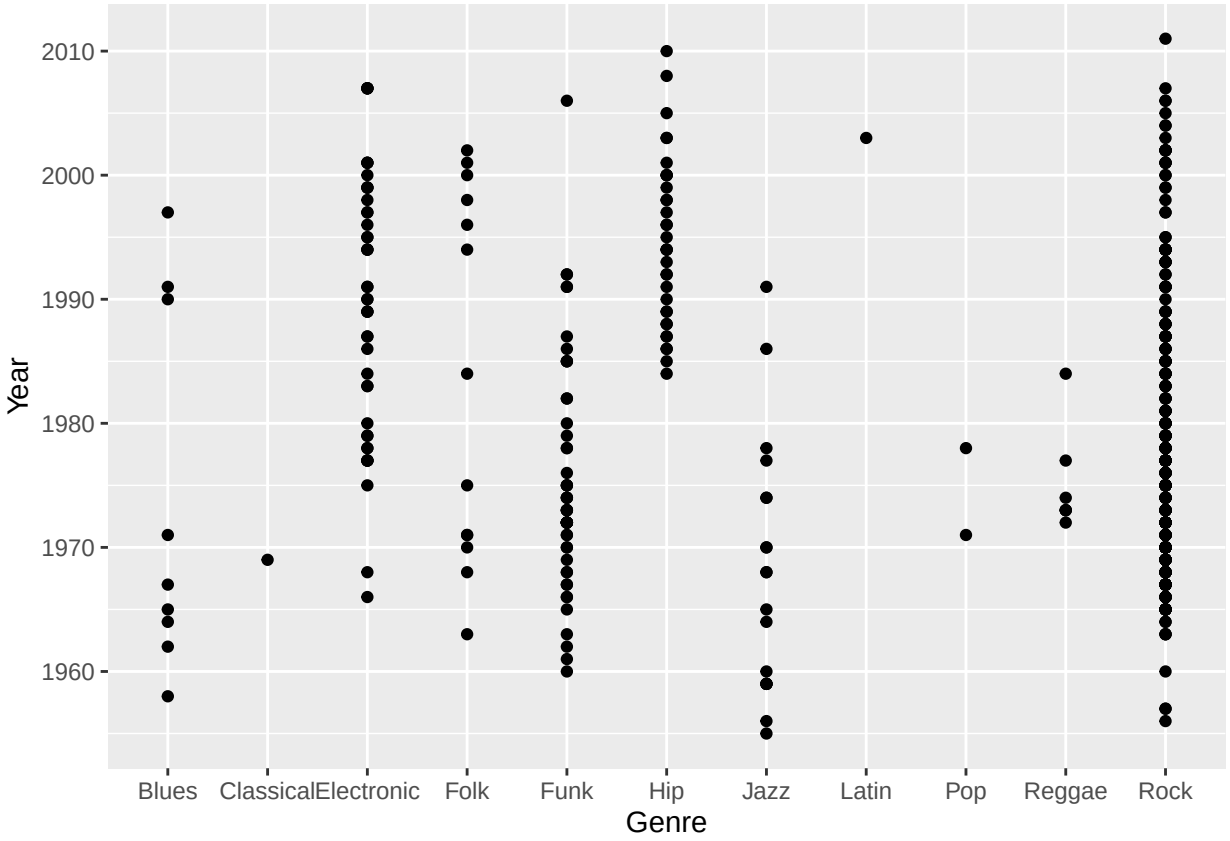

Figure 2: Scatterplot of Genre vs. Year

Although the fit using vglm() can be calculated, when I try to make a binned residual plot, it comes with an error indicating that it might be better for ordinal categorical data rather than nomial. You can find the binned residual plot in the Appendix.

Result

To interpret this model, we have to look at the log risk ratios. We will be using the baseline Genre, Blues, and the denominator. So, we calculate each log risk ratio by the below equations and we interpret it by these:

$log(\frac{P(Y=Genre)}{P(Y=Blues)}) < 0$ then $P(Y = Genre)$ is smaller

$log(\frac{P(Y=Genre)}{P(Y=Blues)}) = 0$ then the probabilities are the same

$log(\frac{P(Y=Genre)}{P(Y=Blues)}) > 0$ then $P(Y = Genre)$ is larger

Below is a table with the count of each genre and their calculated probabilities. The appendix has a table of the log risk ratios and the Discussion section will explain each log risk ratio.

Table 1: Probabilities of Genres

Var1	Freq	Prob
Rock	318	0.636
Funk	51	0.102
Electronic	45	0.090
Hip	34	0.068
Jazz	19	0.038
Folk	13	0.026
Blues	9	0.018
Reggae	7	0.014
Pop	2	0.004
Classical	1	0.002
Latin	1	0.002

Discussion

Below is the same table as what is in the Result section but with an added column of the long risk ratios. Here, we remember that if the log risk ratio is greater than zero, the top probability is larger. If it is equal to zero, the top probability is the same. If it is less than zero, the top probability is smaller.

Table 2: Log Risk Ratios of Genres

Var1	Freq	Prob	Log.Risk.Ratio
Rock	318	0.636	3.5648268
Funk	51	0.102	1.7346011
Electronic	45	0.090	1.6094379
Hip	34	0.068	1.3291359
Jazz	19	0.038	0.7472144
Folk	13	0.026	0.3677248
Blues	9	0.018	0.0000000
Reggae	7	0.014	-0.2513144
Pop	2	0.004	-1.5040774
Classical	1	0.002	-2.1972246
Latin	1	0.002	-2.1972246

From the table above, on one hand, it is clear that the genres of Rock, Funk, Electronic, Hip, Jazz, and Folk are greater than zero, meaning that those probabilities are larger than the baseline of Blues. On the other hand, Reggae, Pop, Classical, and Latin log risk ratios are less than zero, meaning that these probabilities are smaller than the baseline of Blues. As stated before, the Blues genre log risk ratio is equal to zero because it is considered the baseline.

So, to answer the first main question of what will the top three categories be over the next few years, it seems that Rock, Funk, and Electronic have the best chance of being the top album genre in the next few years

based on the log risk ratios but the scatterplot in the Method section shows that Hip Hop could be popular as well. The second question of how will the popularity of the different categories change over time can be determined the most by looking at the density plot in the Introduction section. It seems that in the most recent years, Rock, Hip Hop, Electronic, Funk, and Latin were the highest ranking genres. We see that Latin has only one data point, so it would not be likely that that would be a popular genre in later years. Rock has always been popular and Hip Hop has been popular lately. Electronic is dissipating over the years.

Bibliography

Holtz, Yan. "Help and Inspiration for R Charts." The R Graph Gallery, www.r-graph-gallery.com/index.html.

"Latex Equation Editor." LaTex Equation Editor, www.tutorialspoint.com/latex_equation_editor.htm.

Master of Science in Statistical Practice Presentations.

"Search All 21,141 CRAN, Bioconductor and GitHub Packages." R Documentation and Manuals | R Documentation, www.rdocumentation.org/.

Appendix

Here is a scatterplot of only the top genre per year.

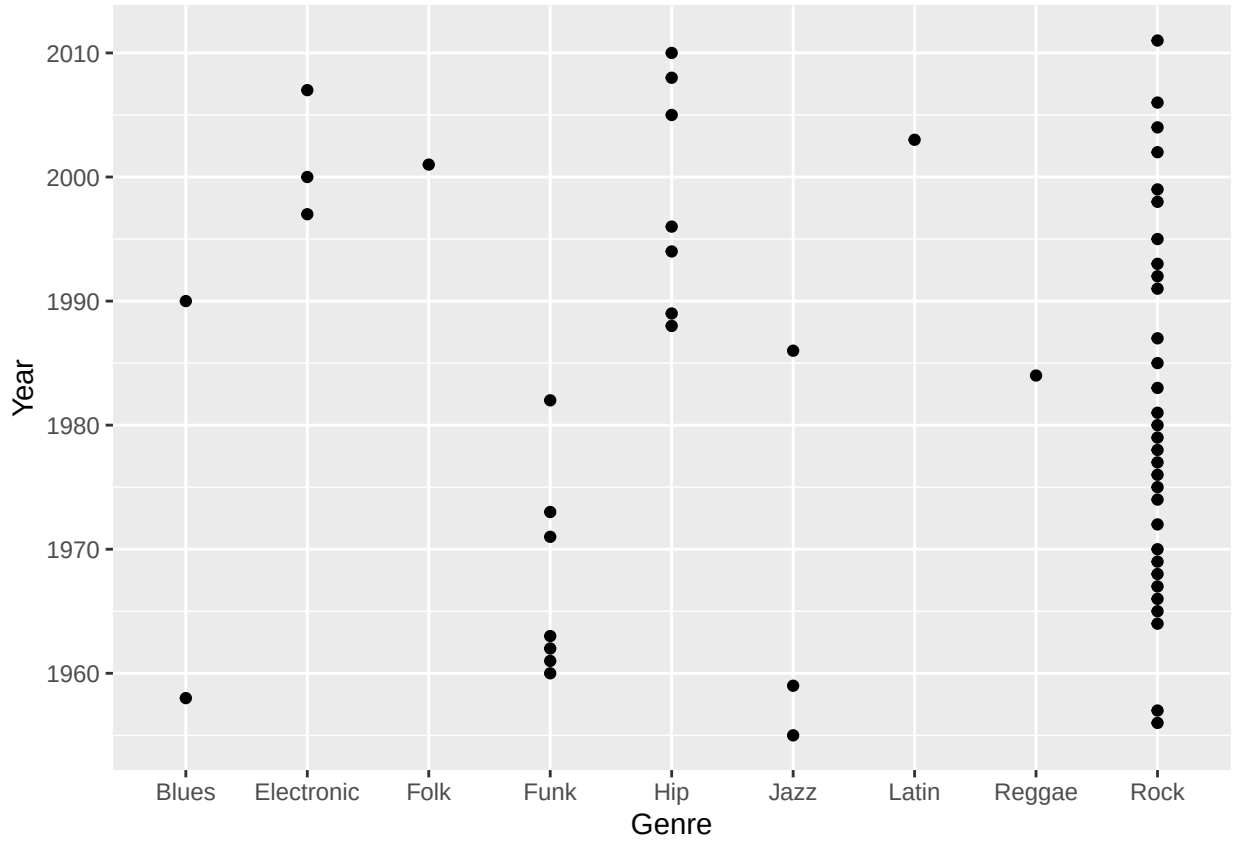

Figure 3: Scatterplot of One Genre per Year

Here are the coefficients of the multinom() regression:

```
## # weights:  33 (20 variable)
```

```
## initial  value 1198.947636
## iter  10 value 688.920445
## iter  20 value 656.472081
## iter  30 value 620.031090
## iter  40 value 601.575931
## iter  50 value 600.707982
## iter  60 value 600.707502
## iter  60 value 600.707500
## final  value 600.707481
## converged
```

Table 3: multinom() Regression Coefficients

	(Intercept)	Year
Classical	7.222475	-0.0047589
Electronic	-145.716668	0.0742203
Folk	-21.760116	0.0111753
Funk	133.400099	-0.0666285
Hip	-203.819703	0.1032464
Jazz	133.135839	-0.0669971
Latin	-13.644621	0.0057905
Pop	7.701669	-0.0046495
Reggae	23.763280	-0.0121335
Rock	40.765879	-0.0187998

Here are the coefficients of the vglm() regression:

Table 4: vglm() Regression Coefficients

	x
(Intercept):1	69.7488152
(Intercept):2	222.5552972
(Intercept):3	-182.1432944
(Intercept):4	-103.1821457
(Intercept):5	42.7793863
(Intercept):6	-265.6830079
(Intercept):7	268.6737653
(Intercept):8	-557.4605013
(Intercept):9	54.1781922
(Intercept):10	41.4995995
Year:1	-0.0371094
Year:2	-0.1157430
Year:3	0.0908456
Year:4	0.0504805
Year:5	-0.0225722
Year:6	0.1326361
Year:7	-0.1376756
Year:8	0.2767881
Year:9	-0.0299842
Year:10	-0.0229297

This plot below shows the amount of times that each genre has been included in the dataset of 500 album genres. It is clear the Rock is the genre that is the most popular over the time span of 56 years (1955-2011) with World and Stage genres not in the dataset at all.

It is clear that this binned residual has a zero percent good fit because of the drastically different categorical data. Also, it really limits what I can do with the data because the data is nomial categorical rather than ordinal.

```
## Warning: Probably bad model fit. Only about 0% of the residuals are inside the error bounds.
```

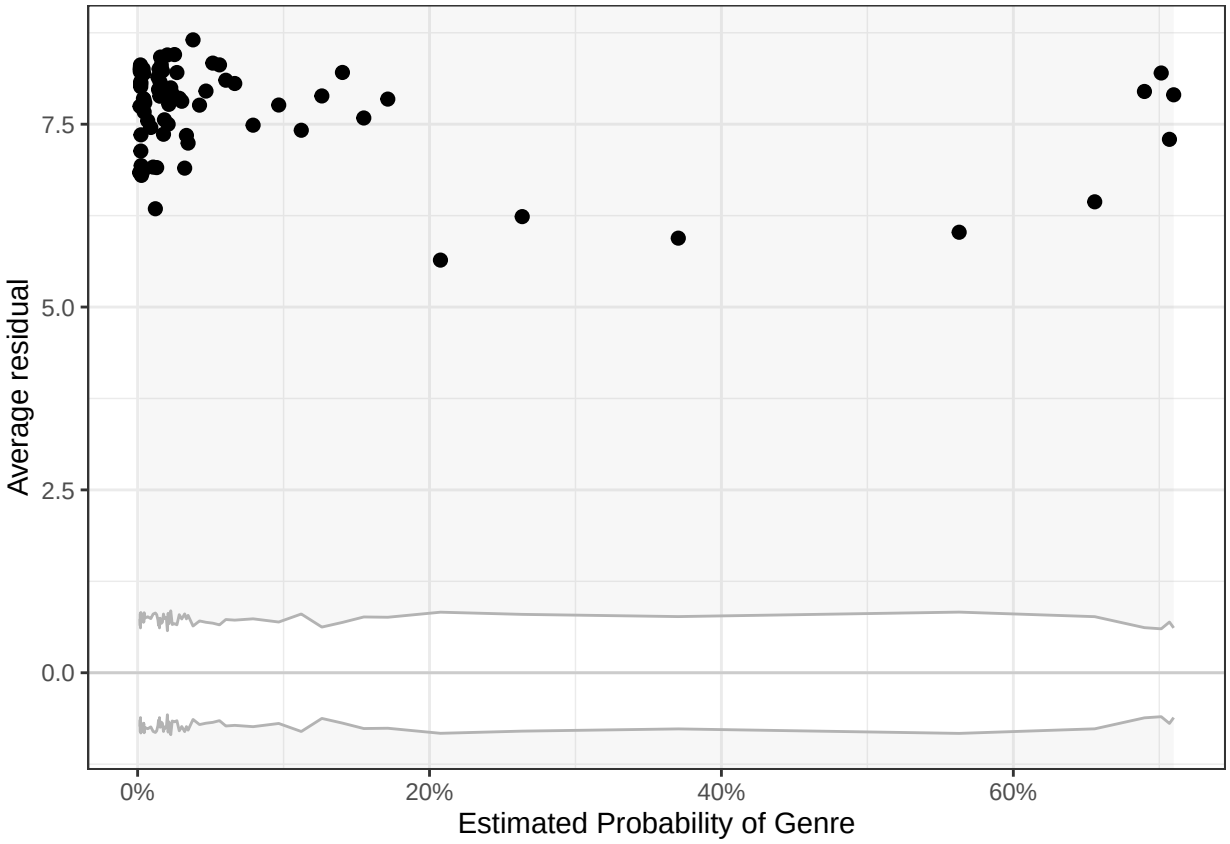

Figure 4: Binned Residual Plot

Here is a bar plot of what I wanted to do with the data first. I wanted to rank each genre based on their level of intensity. So, I ranked them 1-13, where 1 is the lowest intensity and 13 is the highest, and plotted them. I wanted the x axis to have the genres in order from 1-13, but the data is still nomial categorical rather than ordinal. So, it was tedious and ultimately took to long to order that data without doing it manually.

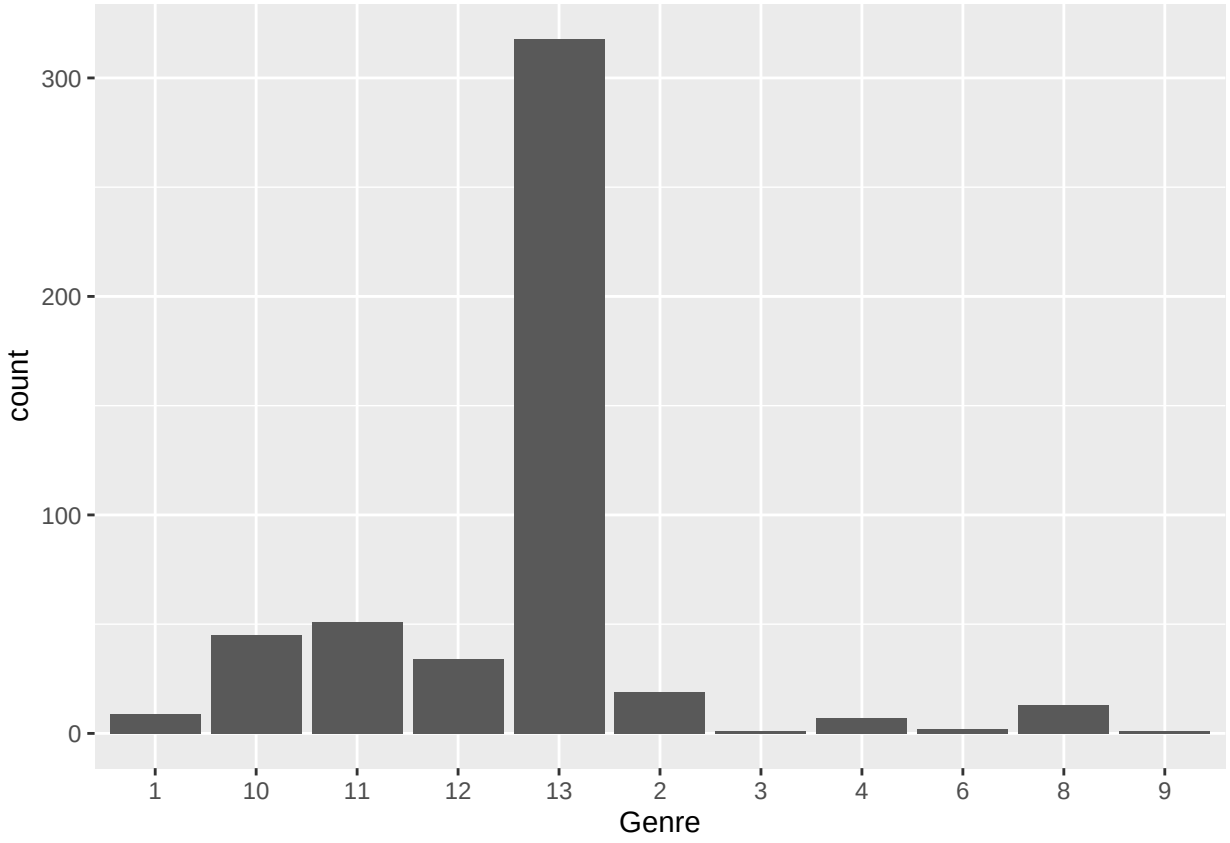

Figure 5: Multinomial Model Coefficients