MET2J Research Project: Initial Report

Bocata de Panceta

23rd of January 2024

1 Introduction

The United States has rich musical diversity that has seen many musicians and genres rise and fall over the years. One of the most essential time periods in terms of the development of music was the 20th Century. Many genres became popular throughout this period, including Jazz, Swing, Rock and Roll, Hip-Hop, and Electronic music. As these musical trends arose, so did the usage of specific instruments. For example, piano is more prevalent in Jazz than in Rock and Roll. As such, this paper will aim to answer the research question: "*are there trends in the frequencies of use of different instruments during the 20th century in the United States?*"

2 Methodology

The data necessary to carry out the analysis was extracted from *'DBPedia'* [1], containing entries for all individuals within the database. We filtered the data down to only musicians (their name), country of origin, year of birth, and instrument played, resulting in 6533 observations. Removing missing values resulted in 2911 observations.

To operationalize the research question, we define the following terms: (i) 'frequency' as the number of 'instrument' players in a given decade; (ii) 'trend' as the changes in the above frequencies per instrument for all instruments included, decade by decade; and (iii) '20th Century' as the ten decades within this century.

These definitions lead to more specific data questions, including: (i) How frequently did specific instruments (Piano, Guitar, Drums, Music software, Voice) occur in each decade in the 20th century? and (ii) how did these frequencies vary across the decades, and what trends did this variation cause (see Figure 1)?

To work with the data, we first had to clean it in Python [2]. We were able to do this by looping through the separate .json files and reading the data in. In Python, utilizing loops made reading this complicated data easier. After filtering the data with 'if' statements to determine whether the desired values were present, we exported to .csv to work with the data and plot in R.

In R, it was extremely difficult to change the values that had very niche names (e.g., ’Fender Stratocaster’ instead of ’Guitar’) into the generic names for more streamlined data. It was easier to remove rows with missing values and then plot the data; however, we will have to spend more time in R over the next few days to perfect our filtering and produce more plots.

3 Results

After plotting our filtered data, trends of instrument use in the United States became noticeable. For example, Figure 1 illustrates a rise in guitar use in mid century with a spike in the 50s, whereas drums use spiked later in the 70s and singing in the 60s-70s. Furthermore, we witness the introduction of music software starting in the late 60’s and its gradual growth until the end of the century.

With a more detail-specific and complete database (that includes relevant but previously removed entries), we will be able to better answer the research question and potentially compare the trends in instruments to the popularity of specific genres over the same time period.

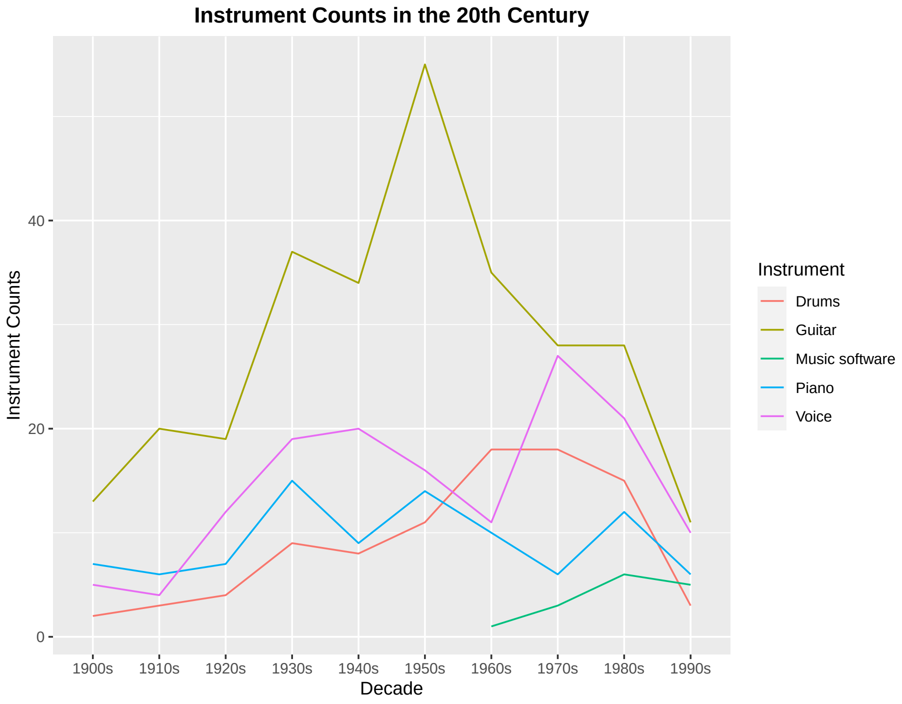

Figure 1: Instrument Counts Per decade in the 20th Century

4 Discussion

From these results, we have already interpreted the meaning of the trends. We have found frequencies for different instruments across the 20th century within the US. This reassures us that our data questions are answerable with our data, but also that we could maybe add additional data questions, such as: How does the variation in instrument prevalence per decade compare to the change in genre popularity over the same time period?

It is relevant to note certain limitations that we are already aware of. The sourced information belongs to English-speaking Wikipedia, which ignores much information from non-English-speaking regions. Furthermore, when selecting musicians from the unfiltered dataset, some were missing a field of information relevant to our data, so their entry was discarded in the data cleaning process (and is unrecoverable in the scope of our project since a combination of fields is necessary).

As for the next steps in data operations, we aim to recover some of the data lost during the preliminary data cleaning (e.g., all musicians with only cities or states as their place of birth were filtered out). As we advance, we aim to find ways to include these observations with US cities or state names as belonging to US 'Country'.

Additionally, we aim to introduce a database that includes genre trends over the same period and geographical area as our filtered data and introduce literature on the prevalence of instruments per genre. This way, our current results can be used to threat a narrative that develops in terms of how instrument frequency in the US converges and diverges with genre evolution over the 20th Century.

Setting out our tasks for tomorrow, we will brainstorm on where to go with the research question, i.e., with the trend we have outlined for the US, what do we think this means in the US music world, and how can we verify that? Two of us will then work on finding the best database and literature to introduce into our project, while the other two go back to Python to improve filtering to recover data and go through the steps in R with the better dataset, to then all keep working with R, trying out and selecting plots, eventually combining the new dataset and external datasets.

References

[1] Jens Lehmann et al. "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia". In: *Semantic web* 6.2 (2015), pp. 167–195.

[2] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.