國立東華大學　應用數學系

碩士論文

指導教授：曹振海　博士

爵士音樂家資料視覺化

Visualization for Jazz Musicians

研究生：呂一昕　撰

中華民國 112 年 6 月

學位考試委員會審定書
Certificate of Approval of Examination Committee

國立東華大學　　應用數學系碩士班

研究生　　呂一昕　　　　　　　　君所提之　論文

National Dong Hwa University
The Thesis Graduate Student Proposed

（題目）　爵士音樂家資料視覺化
Title　　Visualization for Jazz Musicians

經本委員會審查並舉行口試，認為符合　碩士　學位標準。

After evaluation and the oral examination by the committee members, the student complies with the　master　degree

學位考試委員會召集人 The Convener of Examination Committee	*施銘杰*	簽章
委　員 Committee Member	*高竹嵐 Chu-Lan Ku*	簽章
委　員 Committee Member	*曹振海*	簽章
委　員 Committee Member	*施銘杰*	簽章
指導教授 Advising Professor	*曹振海*	簽章
系主任（所長）The Director of Department	*吳韋瑩*	簽章

中華民國　112　年　6　月　13　日

ROC　　　　　Year　2023　Month　Jun　Date　13

國立東華大學
NATIONAL DONG HWA UNIVERSITY

學位論文原創性聲明書
DECLARATION OF THESIS/DISSERTATION ORIGINALITY

學位論文題目： 爵士音樂家資料視覺化

Thesis/Dissertation Title： Visualization for Jazz Musicians

本人在此聲明，所呈交的學位論文是在指導教授曹振海的指導下，由個人獨立研究所完成之最終版本。本人對論文內容負責，除了文中已經標註引用處的內容外，論文不包含任何其他人已經發表或撰寫過的研究成果。對本研究及學位論文做出重要貢獻的個人和組織，均已在文中以明確方式標明。

該論文內容如有違反學術道德或學術規範的行為，如造假、變造、抄襲、研究成果重複發表或未適當引註、以違法或不當手段影響論文審查、不當作者列名等，本人願意承擔由此而產生的法律責任和法律後果。

I declare that the thesis/dissertation herein is the final version of my work, which is composed and accomplished individually under the guidance of my supervisor, Prof. Chen-Hai Tsao. I am responsible for the contents of this thesis/dissertation: It contains no research result that was previously published or written by another person. Information derived from published and unpublished work of others has been acknowledged in the text, and a list of references is given . Any contribution made by other individual or organization is explicitly acknowledged in the thesis/dissertation.

If any research misconduct, including fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results, is discovered in my thesis/dissertation, I am willing to bear corresponding legal responsibilities and all the results therefrom.

聲明人 Declarant： 呂一昕

日期 Date： 2023/07/11

(yyyy/mm/dd)

致謝詞

　　研究所的這段時間對我而言雖然是漫長的兩年多，但在上帝所創造的歷史長河中，如螻蟻一般渺小，光陰似箭、歲月如梭，終於完成學位論文。回想起當時毅然決然走進研究所，就如同昨日一般歷歷在目，中研院工作就像剛結束的任務，也因著這些經驗引發我對統計、資料分析和研究的興趣。我很幸運地回到最熟悉的地方繼續走上進修之旅，也很開心能跟著曹振海教授共同觀察著音樂和統計所擦出的火花，最後才有這篇論文的產生。

　　首先特別感謝我的父母，沒有他們支持，我沒有辦法健康地成長到現在，無論是財務、心理或是其他方面，都給了我最大的支持與鼓勵，我的爸爸常跟我說：「最重要的就是訓練解決問題的能力，大多數的研究生甚至是博士班都不一定有這樣的能力。」感謝我的爸爸給我很多心理和技術上的支持，所以我常常開玩笑說，我一個禮拜都要開兩次會。

　　感謝我的指導老師 曹振海 教授，老師給了我非常多自由發揮的空間，讓我能在眾多的研究生中自由翱翔，肆意展現我的才華和智慧。很謝謝老師在我研究所階段不厭其煩地幫助我、提醒我什麼事情才是最重要的，也讓我周末有娛樂時間去好好放鬆、運動、彈琴。在老師所指導的時間內，不斷刷新我對這個領域的認知，原來研究所也可以這麼好玩。在學術上給予我非常大的支持和協助、在我有興趣的內容上給我無窮無盡的資源分享、在我需要的地方上提供很多的新創意，在這個階段所獲得的東西遠超過我所求所想，真的再次獻上誠摯的感謝。

　　感謝音樂系 魏廣晧 教授，感謝老師在爵士音樂上給予最專業的建議和資訊的參考，結合在我的研究當中。當時也是因著老師的建議才有現在的成果，討論之中使我更了解爵士音樂中還有這麼多的發展與故事。

　　感謝我的女朋友 黃筱嵐，陪在我身邊用自身的努力激勵我持續努力，也在我心情沮喪時給我最大的關心。因為自身能力懷疑時，給我信心、不斷鼓勵我，讓我度過這段研究所比較辛苦的時期。

　　感謝所有協助過我的人，因為有你們才有現在的我，如果致謝詞中沒提到，我深感抱歉，篇幅有限，所以在此獻上最誠摯的歉意。

研究生　呂一昕　謹致

摘要

　　我們使用爵士音樂家的資料建立一些視覺化圖形，這些圖能顯示音樂家在樂器、曲風和活躍年代上的相似程度。針對爵士音樂家的資料視覺化，選擇以 Wynton Marsalis 和 Roy Hargrove 為首的 229 位音樂家，並以他們的樂器、曲風和活躍年代做為資料矩陣的變數。我們將這些變數分別處理、運算得出各自的相似度矩陣，再將這些矩陣合併，用 PCA 和 t-SNE 降至二維。把二維的資料點繪製在平面上得到目標圖形，這些圖形比現有的 Linked Jazz 能顯示出更多的資訊。當我們用變數上色時，發現大部分的音樂家都不只演奏單一樂器或曲風。針對樂器和曲風這類多值變數 (multi-value variable)，我們提出以三原色分別代表三種類型、接著依照比例進行調色，以此作為上色標準。

關鍵字：爵士音樂家、資料視覺化、三原色、多值變數

Abstract

We created jazz musician maps that displayed the similarities between musicians in terms of instruments, genres, and active years. For jazz musician data visualization, we selected 229 musicians associated with Wynton Marsalis and Roy Hargrove and used their instruments, genres, and active years as variables. We separated the data matrix by variable and turned each into an affinity matrix, and then we combined all of the affinity matrices and utilized PCA and t-SNE to reduce the matrix to two dimensions, resulting in jazz musician maps that were more informative than Linked Jazz. When we colored by variables, the instruments and genres were multi-value variables, meaning the musicians played more than one instrument or genre. Therefore, we presented that utilized the three primary colors of the CMYK color system to represent three groups and then tuned the color based on the number of each group to be the progressing method of coloring.
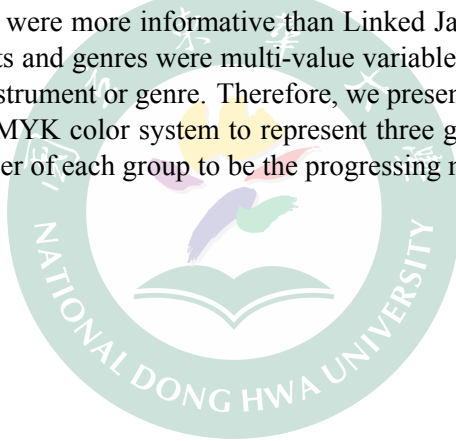
Keywords: jazz musician maps, visualization, multi-value variables

Contents

Chapter 1

Introduction

1.1 Data Visualization

Data visualization is a widely-used technique for presenting information derived from data. According to Chen *et al.* (2007), good visualization should possess certain qualities. Several studies, such as those by Aparicio and Costa (2015), Sadiku *et al.* (2016), and Unwin (2020), discuss the importance of visualization. Furthermore, Ferrari and Russo (2016) introduces Microsoft Power BI, a tool that enables users to create visualizations easily. In a recent TED talk, Bacallado (2020) discusses how data visualization can improve our lives. Given these factors, data visualization is currently a popular and effective information presentation method.

Data visualization is a vast topic that encompasses different kinds and purposes. In the realm of data analytics, scatter plots, histograms, box plots, curves, and network graphs are commonly used. Ahn *et al.* (2011) is an excellent example of data visualization using a network graph to illustrate the relationship between flavors and foods.

To demonstrate a visualization of musicians, *Music Map* and *Linked Jazz* are great graphs. Especially, *Linked Jazz* offers an excellent map with many benefits and options, which includes nearly every jazz musician. The Linked Jazz visualization (Figure 1.1) allows the user to pause the cursor on a musician to view their profile and associated musicians. It provides four modes: "Fixed" (the primary map), "Similar" (connections based on similarity), "Gender" (coloring male and female musicians differently), and "Dynamic" (allowing users to choose which musicians to display). The "Dynamic" mode is particularly noteworthy as it is useful in displaying the social network of selected musicians. For instance, when we select Wynton Marsalis and Roy Hargrove, the map displays both of their associated musicians. However, this network graph leaves room for improvement.

We decide to utilize statistical methods to create a graph that displays jazz musicians and is used as a map. On the map, the musicians are located according to their relationships and similarities. Furthermore, we color and give the weight by instruments, genres, and active years to construct the maps. For example, in the instruments, musicians with similar instruments have a close distance on the map, and they are also in the same hue of colors. Likewise, there is the same representation for other features.

1.2 Jazz Musicians

To facilitate our discussion and visualization, we give a short review of selected jazz musicians. As suggested by Wei (2022), we start with Wynton Marsalis and Roy Hargrove and then collect related musicians in a 2-degree circle sense. Wynton Marsalis and Roy Hargrove are the most important musicians, and Marsalis family members are also essential. In addition to these musicians, we have selected a few other famous and influential musicians for our introduction, referring to 2.2 for the detail of the jazz musicians list. The interested readers are referred to *Music Brainz*.

Wynton Marsalis

Wynton Learson Marsalis is a trumpeter, composer, and educator born on October 18, 1961, in New Orleans, Louisiana. He began playing trumpet at the age of six and in 1980, he joined Jazz Messengers led by Art Blakey. Around 1982, he recorded with Blakey and his first solo album, showcasing his exceptional talent. Marsalis has won nine Grammy Awards and has become a world-renowned musician. According to the Public Broadcasting Service (PBS):
"He is the only artist in any genre to have won GRAMMY Awards in five consecutive years (1983–1987) and the first jazz artist to be awarded the Pulitzer Prize in music. In 1987, Wynton helped launch the Classical Jazz summer concert series at Lincoln Center in New York City." (The post *Wynton Marsalis Biography* from PBS)
Overall, he is considered one of the greatest jazz musicians of all time and has *Wynton Marsalis Official Website*.

Branford Marsalis

Branford Marsalis is a saxophonist, composer, and bandleader born on August 26, 1960, in Breaux Bridge, Louisiana. Branford began playing the clarinet at a young age and later switched to the soprano and tenor saxophones. In 1980, he joined Art Blakey's Jazz Messengers with his brother Wynton Marsalis. He then went on to play with numerous other jazz legends, including Miles Davis and Dizzy Gillespie.

Jason Marsalis

Jason Marsalis is a jazz drummer, percussionist, and also the youngest of the four Marsalis brothers, he was born on March 4, 1977, in New Orleans, Louisiana.

Ellis Marsalis Jr.

Ellis Marsalis Jr. is a pianist, educator, and patriarch of the Marsalis family born on November 14, 1934, in New Orleans, Louisiana. He taught music at the New Orleans Center for Creative Arts and the University of New Orleans, and Many of his students went on to become successful jazz musicians in their own right, including his four sons, Wynton, Branford, Delfeayo, and Jason Marsalis.

Roy Hargrove

Roy Hargrove, born on October 16, 1969, in Waco, Texas, was a renowned trumpeter, composer, and bandleader who grew up in a musical family. He played trumpet and flugelhorn and occasionally served as a vocalist. He received guidance and mentorship from jazz trumpeter Wynton Marsalis. In 1990, he gained worldwide attention after performing with David Murray's band, following which he went on to lead his group and collaborate with jazz legends such as Dizzy Gillespie. Sadly, he passed away in 2018.

Art Blakey

Art Blakey (1919-1990) was an American jazz drummer and bandleader born in Pittsburgh, Pennsylvania, and began playing drums at a young age.
"Art Blakey and the Jazz Messengers was one of the most enduring, popular, reliable, and vital small bands in modern jazz history." (Goldsher (2002): "Hard Bop Academy")

J. J. Johnson

J. J. Johnson (1924-2001) was an American jazz trombonist and composer born in Indianapolis, Indiana. His career spanned over five decades, and he played with many of the most important jazz musicians of his time, including Charlie Parker, Dizzy Gillespie, and Miles Davis. He recorded over 50 albums as a bandleader, and he won numerous awards and accolades for his contributions to jazz. His music continues to be celebrated and studied by jazz musicians and fans around the world.

Clark Terry

Clark Terry (1920-2015) was an American jazz trumpeter, flugelhorn player, and educator born in St. Louis, Missouri. He is perhaps best known for his work as a member of Duke Ellington's orchestra in the 1950s and 1960s, as well as his prolific career as a solo

artist and bandleader. He was also an important educator, and he mentored many young jazz musicians over the course of his career.

Dizzy Gillespie

Dizzy Gillespie (1917-1993) was an American jazz trumpeter, bandleader, and composer born in Cheraw, South Carolina, and grew up in New York City. He was one of the leading figures of the bebop movement in the 1940s, and his virtuosic playing and innovative compositions had a major impact on the development of jazz. He was also an important advocate for Afro-Cuban jazz, and he collaborated with many Latin American musicians over the course of his career.

Miles Davis

Miles Davis (1926-1991) was an American jazz trumpeter, bandleader, and composer born in Alton, Illinois, and grew up in East St. Louis. He was one of the most influential musicians in jazz history, and his innovative playing and groundbreaking recordings had a major impact on the development of music. Davis played with many of the leading jazz musicians of his time, and he collaborated with a wide range of musicians from other genres as well. He was also an important innovator in jazz fusion, and his work in that genre helped to shape the sound of popular music in the 1970s and beyond.

1.3   Chapter Overview

The rest of the thesis is organized as follows: Chapter 2 shares data and its collection, Chapter 3 introduces the methods from data to maps, Chapter 4 displays the result and discusses observation on the maps, and the conclusion is presented in Chapter 5.
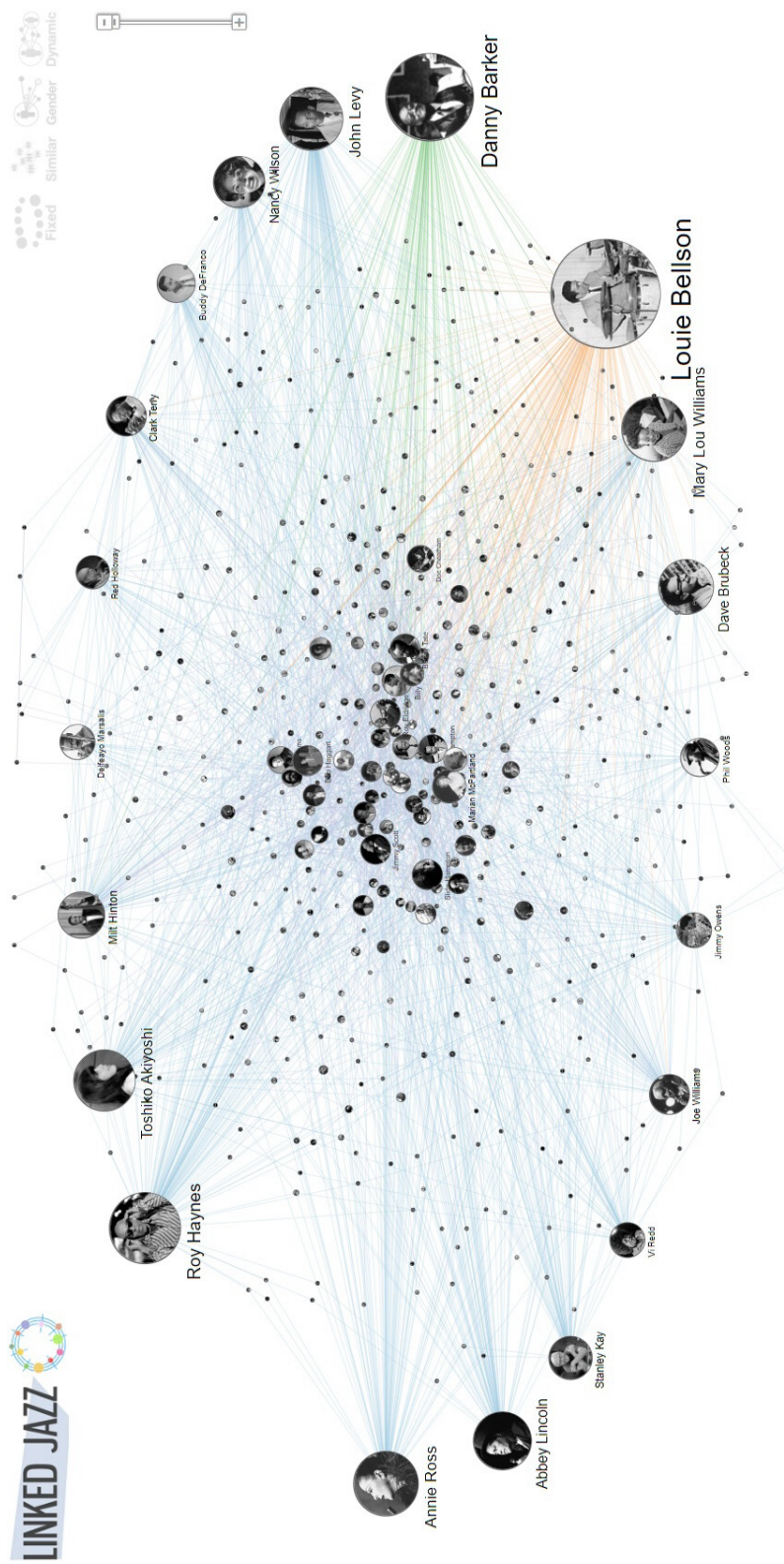
Figure 1.1: Linked Jazz Network Graph

7

Chapter 2

Data

2.1 Original Data Table

Our data consists of 229 jazz musicians and their features: instruments, genres, the start of active years, and the end of active years. (see an example in Table 2.1) We collect the data from the same DBpedia database as *Linked Jazz*.

2.2 Collection

To make our project manageable and meaningful, we restrict to those musicians within the "2-degree" circles of Wynton Marsalis and Roy Hargrove, as suggested by Wei (2022). Specifically for example in Table 2.2, the musicians in "is associatedMusicalArtist of" are in the "1-degree" circle, and they are Branford Marsalis, Clark Terry, Art Blakely, and others. Respectively, musicians of "2-degree" circle can be collected. And this results in our musician name list.

We select instruments, genres, and active years as the primary features for our analysis and use a Python script to extract data from websites, referring to Lu (2022) for the source code.

Table 2.1: Exploring Four Essential Features of Musicians

	Instruments	Genres	Start of active years	End of active years
Wynton_Marsalis	Trumpeter, Composer	Dixieland, Jazz, Classical_music	1980	2022
Branford_Marsalis	Saxophonist, Composer	Jazz	1980	2022
Jason_Marsalis	Drummer, Vibraphonist	Jazz		
Ellis_Marsalis_Jr.	Pianist	Jazz, Classical_music	1949	2020
Roy_Hargrove	Trumpeter, Flugelhornist, Vocalist	Jazz, Latin jazz, M_Base, Soul	1987	2018
Art_Blakey	Drummer, Percussionist	Jazz, Bebop, Hard_bop	1942	1990
J_J_Johnson	Trombonist	Jazz, Bebop, Third_stream, Hard_bop	1942	1996
Clark_Terry	Flugelhornist, Vocalist, Trumpeter	Jazz, Bebop, Hard_bop, Swing_(jazz_performance_style)	1940	2015
Dizzy_Gillespie	Pianist, Vocalist, Trumpeter	Afro_Cuban_jazz, Jazz, Bebop	1935	1993
Miles_Davis	Pianist, Trumpeter, Cornetist, Electric organist, Flugelhornist	Jazz, Modal_jazz, Cool_jazz, Hard_bop, Jazz_fusion	1944	1975

Table 2.2: *DBpedia Page for Wynton Marsalis*

Property	Value
Instruments	Trumpeter, Composer
Genres	Dixieland, Jazz, Classical_music
Start of Active years	1980
End of active years	Present
is associatedMusicalArtist of	Branford_Marsalis Clark_Terry Art_Blakely ⋮
⋮	⋮
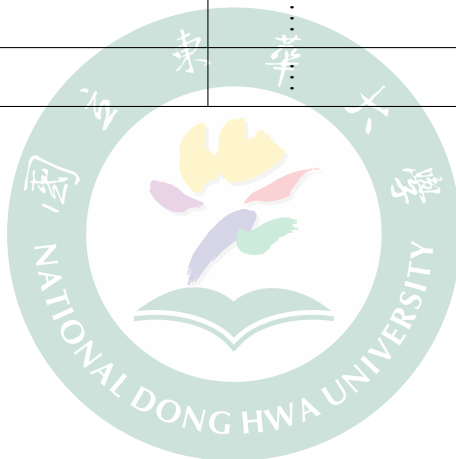
Chapter 3

Methods

This chapter introduces the methods used to transform the original data table into jazz musician maps, referring to Figure 3.3 for the flowchart. Section 3.1 and 3.2 construct the matrix based on the pairwise similarity between musicians. We then use dimension reduction as a visualization method to construct the maps. In the last section, we discuss and present how we color the data points (musicians) by the multi-value variables (instruments and genres).

3.1 Working Data Matrices

For easy exposition, we take Wynton Marsalis and Roy Hargrove as an example. As shown in Table 3.1, we decide to separate those features and handle them independently.

Table 3.1: Original Data Table

	Instruments	Genres	Start year	End year
Wynton Marsalis	composer, trumpeter	Dixieland, Jazz, Classical_music	1980	present
Roy Hargrove	composer, trumpeter, vocalist, hornist	Jazz, Latin jazz, M_Base, Soul	1987	2018

3.1.1 Instruments and Genres

We collect all of the major instruments to create Table 3.2. Likewise, the genres are used the same processing method.

Table 3.2: Instruments

	composer	trumpeter	vocalist	hornist
Wynton Marsalis	1	1	0	0
Roy Hargrove	1	1	1	1

3.1.2 Active Years

The active years are in the third and fourth columns of Table 3.1. We utilize one-hot encoding and fill in the years of their careers to create Table 3.3.

Table 3.3: Active Decades

	1970	1980	1990	2000	2010	2020
Wynton Marsalis	0	1	1	1	1	1
Roy Hargrove	0	1	1	1	1	0

We create another two features: the number of active decades and the middle of active years. (Table 3.4 and Table 3.5)

Table 3.4: Number of Active Decades

	0	1	2	3	4	5	6
Wynton Marsalis	0	0	0	0	0	1	0
Roy Hargrove	0	0	0	0	1	0	0

Table 3.5: Middle of Active Years

	1990	2000
Wynton Marsalis	0	1
Roy Hargrove	1	0

3.2 Affinity Matrices

We consider using Jaccard.c to construct the affinity matrices.

3.2.1 Jaccard

Jaccard is a technique for computing the similarity between two vectors (or the rows in the data matrix). Since we have the similarity value, we then compute the affinity matrix.

Assume that $x = (x_1, \cdots, x_n)$, $y = (y_1, \cdots, y_n)$, Jaccard similarity excludes "0-items" ($x_i = y_i = 0$) and is defined by:

$$Jaccard(x, y) = \frac{|x \cap y|}{|x \cup y|} \tag{3.1}$$
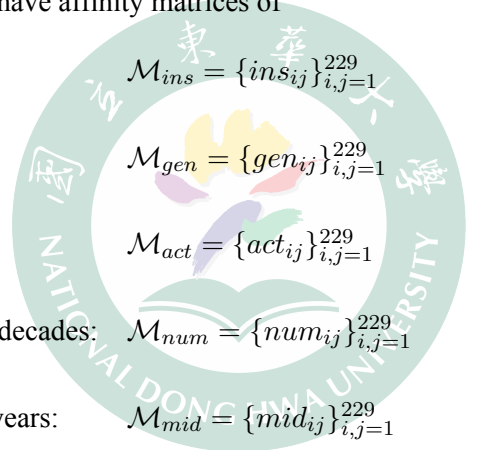
3.2.2 Jaccard.c

Tsao *et al.* (2023) suggests a new variation of Jaccard that retains some of the "0-items" because sometimes they are meaningful to the similarities. Therefore, we add the constant in the formula:

$$Jaccard.c(x,y) = \frac{|x \cap y| + c}{|x \cup y| + c}, \ \forall \, c > 0. \tag{3.2}$$

We compare the original Jaccard with Jaccard.c in Appendix A and notice that Jaccard.c is better than the original one.

3.3 Combinations

To recapitulate, we have affinity matrices of

$$
\begin{cases}
\text{instruments:} & \mathcal{M}_{ins} = \{ins_{ij}\}_{i,j=1}^{229} \\[2mm]
\text{genres:} & \mathcal{M}_{gen} = \{gen_{ij}\}_{i,j=1}^{229} \\[2mm]
\text{active years:} & \mathcal{M}_{act} = \{act_{ij}\}_{i,j=1}^{229} \\[2mm]
\text{number of active decades:} & \mathcal{M}_{num} = \{num_{ij}\}_{i,j=1}^{229} \\[2mm]
\text{middle of active years:} & \mathcal{M}_{mid} = \{mid_{ij}\}_{i,j=1}^{229}
\end{cases}
$$

We combine those affinity matrices in (3.3) to construct the mixed affinity matrix ($\mathcal{M}(all) = \{\mathcal{M}(all)_{ij}\}_{i,j=1}^{229}$) because this combination consists of pairwise similarity with all features.

$$\mathcal{M}(all)_{ij} = \sqrt{\frac{1}{5}\Big((ins_{ij})^2 + (gen_{ij})^2 + (act_{ij})^2 + (num_{ij})^2 + (mid_{ij})^2\Big)} \tag{3.3}$$

When we visualize $\mathcal{M}(all)$, we surmise the single-value variables (active years) are strong information to affect the map. For the balance, we give the active years features lower weights. (3.4)

$$\mathcal{M}(all)_{ij}^* = \sqrt{\frac{1}{3}\Big((ins_{ij})^2 + (gen_{ij})^2 + 0.4(act_{ij})^2 + 0.3(num_{ij})^2 + 0.3(mid_{ij})^2\Big)} \tag{3.4}$$

Following the weight tuning tests, we derive formulas for computing the mixed affinity matrix, where $\mathcal{M}(\cdot)_{ij}$ represents the i-th row and j-th column member in $\mathcal{M}(\cdot)$.

Instruments:

$$\mathcal{M}(ins)_{ij} = \sqrt{\frac{1}{10}\Big(8(ins_{ij})^2 + (gen_{ij})^2 + 0.4(act_{ij})^2 + 0.3(num_{ij})^2 + 0.3(mid_{ij})^2\Big)}$$

Genres:

$$\mathcal{M}(gen)_{ij} = \sqrt{\frac{1}{10}\Big((ins_{ij})^2 + 8(gen_{ij})^2 + 0.4(act_{ij})^2 + 0.3(num_{ij})^2 + 0.3(mid_{ij})^2\Big)}$$

Number of active decades:

$$\mathcal{M}(num)_{ij} = \sqrt{\frac{1}{3.7}\Big((ins_{ij})^2 + (gen_{ij})^2 + 0.4(act_{ij})^2 + (num_{ij})^2 + 0.3(mid_{ij})^2\Big)}$$

Middle of active years:

$$\mathcal{M}(mid)_{ij} = \sqrt{\frac{1}{3.7}\Big((ins_{ij})^2 + (gen_{ij})^2 + 0.4(act_{ij})^2 + 0.3(num_{ij})^2 + (mid_{ij})^2\Big)}$$

3.4 Dimension Reduction

Dimension reduction is proposed as a visualization method since reducing high-dimensional data to two dimensions.

3.4.1 PCA

The PCA is a conventional method for reducing dimensionality in multivariate analysis, referring to Härdle, Simar, *et al.* (2007) for more detail.

3.4.2 t-SNE

Van Der Maaten and Hinton (2008) presents t-SNE, a technique for compressing high-dimensional data matrices into two or three dimensions, and then visualizing them.

Assuming that $\mathcal{X}$ is the data matrix and $\mathcal{X} = (x_1, x_2, \cdots, x_n)^T$, where $x_i \in \mathbb{R}^p$ for all $i = 1, \cdots, n$, the conditional probability of x_i being considered a neighbor of x_j is denoted by $p_{j|i}$.

$$p_{j|i} = \frac{exp(-\|x_i - x_j\| / 2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_i - x_k\| / 2\sigma_i^2)} \tag{3.5}$$

where σ_i^2 is the variance of centering on x_i.

$$P_i = (p_{1|i}, \cdots, p_{n|i}), \forall i = 1, \cdots, n.$$

Defined by:

$$P = (P_1, \cdots, P_n)^T.$$

The t-SNE randomly selects n data points ($x_i'; \forall i = 1, \cdots, n$) in $\mathbb{R}^2$ by sampling from a Gaussian distribution. It then transforms their probabilities using a t-distribution based on the Euclidean distances between the points as follows:

$$q_{ij} = \frac{(1 + \left\|x_i' - x_j'\right\|^2)^{-1}}{\sum_{k \neq i}(1 + \left\|x_i' - x_k'\right\|^2)^{-1}} \tag{3.6}$$

To construct a similar system as P:

$$Q_i = (q_{ij}, \cdots, q_{ij}), \forall i = 1, \cdots, n.$$

Defined by:

$$Q = (Q_1, \cdots, Q_n)^T.$$

The KL divergence is utilized to measure the difference between P and Q as two probability distributions. The loss function (C) is given as follows:

$$C = \sum_i KL(P_i\|Q_i) = \sum_i \sum_j p_{j|i} log\frac{p_{j|i}}{q_{ij}} \tag{3.7}$$

The t-SNE uses gradient descent to minimize the loss function (C), and we typically choose to construct the data matrix $\mathcal{X}'$ and $\mathcal{X}' = (x_1', x_2', \cdots, x_n')^T$. As a result, we illustrate the map by x_i' since $x_i' \in \mathbb{R}^2$.

3.5 Coloring

Meanable coloring for data points improves the map specifically. In our working data matrices, instruments and genres are multi-value variables, and this situation leads to difficulty with coloring figures. We utilize the three primary colors (cyan, magenta, and yellow) of the CMYK color system to indicate three groups of variables, we then color the data points by tuning the rate of three primary colors.

3.5.1 Grouping

We use PCA to compress the data into lower dimensions such as a text mining technique that is to turn the tokens into topics or themes. For example, the working data matrix is musicians by instruments. However, we utilize a data matrix (instruments by musicians) to obtain the relationship between instruments.

Instruments

In Figure 3.1, the cyan words indicate accompaniment instruments, while the drummer and percussionist are positioned similarly in a band. Bassists and keyboardists are frequently used as the foundation of jazz.

The magenta words in the fourth quadrant include most wind instruments, a pianist, and a composer. In our musician list, some of them are both composers and trumpeters, and Wynton Marsalis, Roy Hargrove, Dizzy Gillespie, and Miles Davis are classical examples.

The instruments in yellow whose PC1 axis is less than -0.2. In this area, songwriters, singers, and vocalists have a closely linked relationship. It is important to note that the distinction between vocalists and singers is their relationship with the band. Vocalists typically sang as part of a band, while singers can perform as soloists or in a group. It is interesting to note that most producers are formerly guitarists in our dataset.

Furthermore, PC1 and PC2 explain 34.55% of the variation. While it seems not a high value, it is good enough and reasonable to group those instruments.

Table 3.6: Instruments

Percussions	drummer, bassist, percussionist, keyboardist
Winds	pianist, composer, saxophonist, flutist, trumpeter, trombonist, hornist
Vocals	vocalist, singer, songwriter, guitarist, producer

Referring to 3.6 and Figure 3.1, it should be noted that composer, producer, and songwriter are not traditionally considered instruments, although they are listed as such in DBpedia. In our dataset, instruments encompass not only variables or features but also positions or roles within a jazz band. Additionally, we observed that the piano and keyboard are distinct instruments with different roles within a jazz band. Pianists often take on solo performances, while keyboardists typically provide accompaniment. Similarly, the terms "singer" and "vocalist" represent different roles. For instance, John Mayer is a renowned singer who performs solo within a band. However, determining whether to treat these roles as distinct instruments is a subjective decision, as there is no definitive answer.

Genres

In Figure 3.2, this PCA map explains 28.93% of the variation, and we notice a few genres staying in the same area. Blue, soul, and neo-soul are similar, and the left side of the

second quadrant is the "Bop" style, including post-bop, bebop, and hard bop. Moreover, it is reasonable to classify the genres whose value of PC1 is less than -0.1 into Genre 1, as those genres are much different from the others in our dataset, and Genres 2 and 3 both are clearly grouped on the map.

Table 3.7: Genres

Genre 1	hard bop, bebop, post-bop, modal-jazz, fusion, third stream, avantgarden, free improvisation, free jazz
Genre 2	cool jazz, mainstream jazz, afro cuban jazz, west coast jazz, swing, big band, Latin jazz
Genre 3	blues, funk, soul, neo-soul, gospel

3.5.2   Tuning

For easy exposition, we take a few musicians and their instruments as an example. (see Table 3.8)

Table 3.8: Instruments

	drummer	bassist	composer	trumpeter	horn	pianist	vocalist	guitarist
Wynton Marsalis	0	0	1	1	0	0	0	0
Roy Hargrove	0	0	1	1	1	0	1	0
Miles Davis	1	1	1	1	1	1	0	1
Art Blakey	1	0	0	0	0	0	0	0

We then turn instruments into groups and construct Table 3.9.

Table 3.9: Instruments Grouping

	Percussions	Winds	Vocals
Wynton Marsalis	0	2	0
Roy Hargrove	0	3	1
Miles Davis	2	4	1
Art Blakey	1	0	0

In Table 3.10, we calculate the rate of three groups as the percentage of colors.

Table 3.10: Three Primary Colors

	Cyan	Magenta	Yellow
Wynton Marsalis	0	100%	0
Roy Hargrove	0	75%	25%
Miles Davis	28.6%	57.1%	14.3%
Art Blakey	100%	0	0

Now we have the colors for each musician in Table 3.11.

Table 3.11: Colors

	Colors
Wynton Marsalis	
Roy Hargrove	
Miles Davis	
Art Blakey	
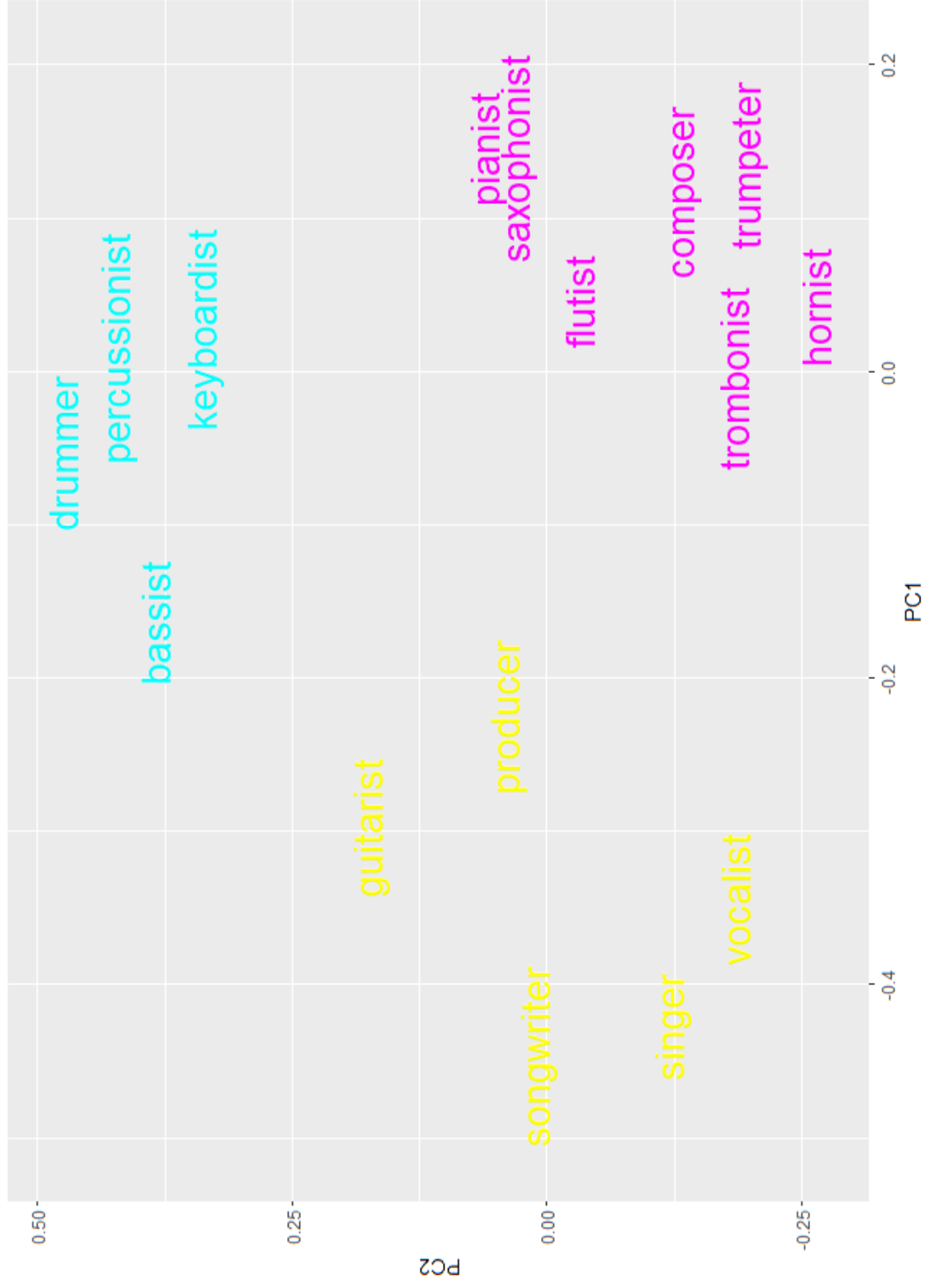
Figure 3.1: PCA for instruments
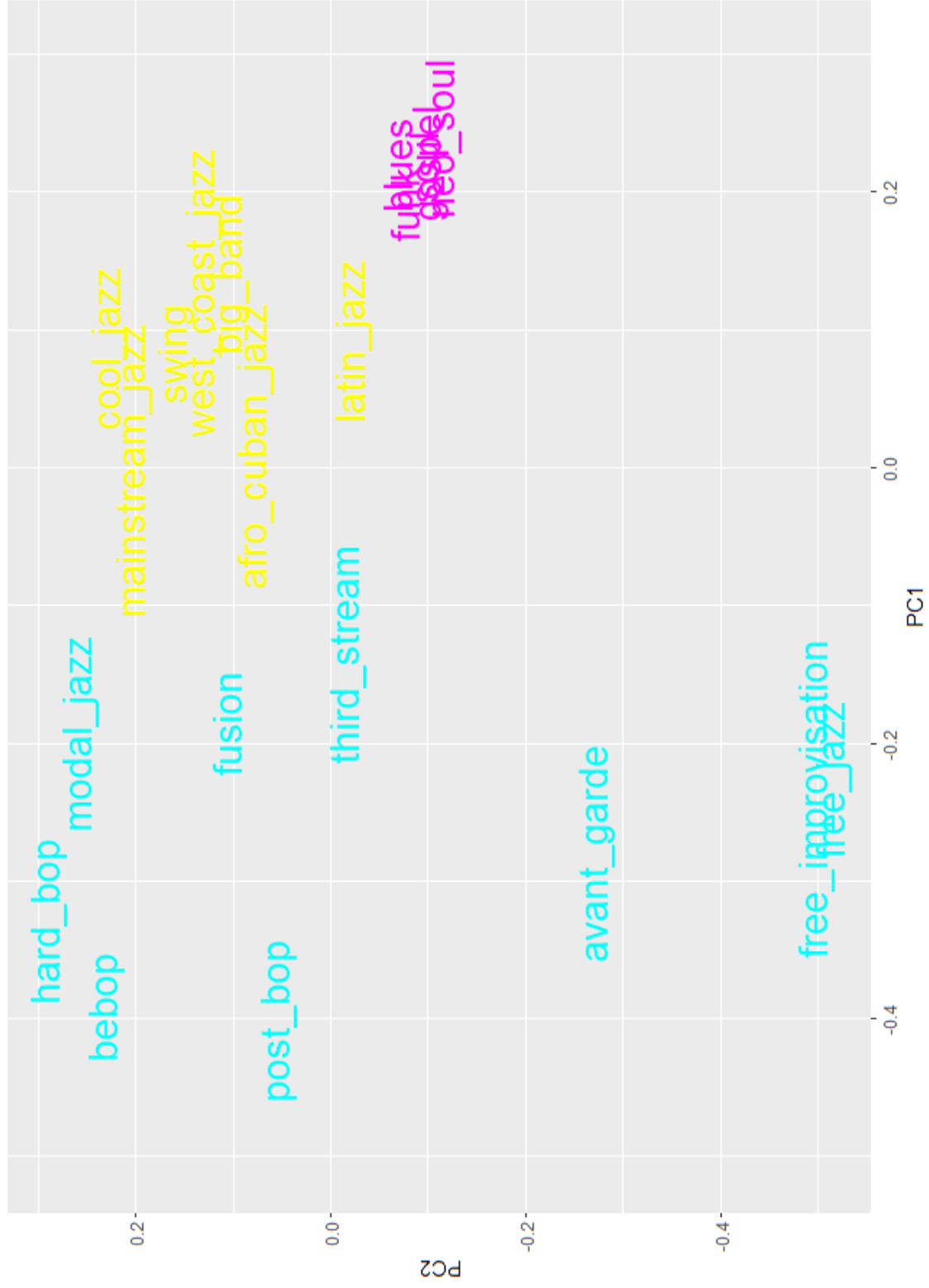
Figure 3.2: PCA for genres
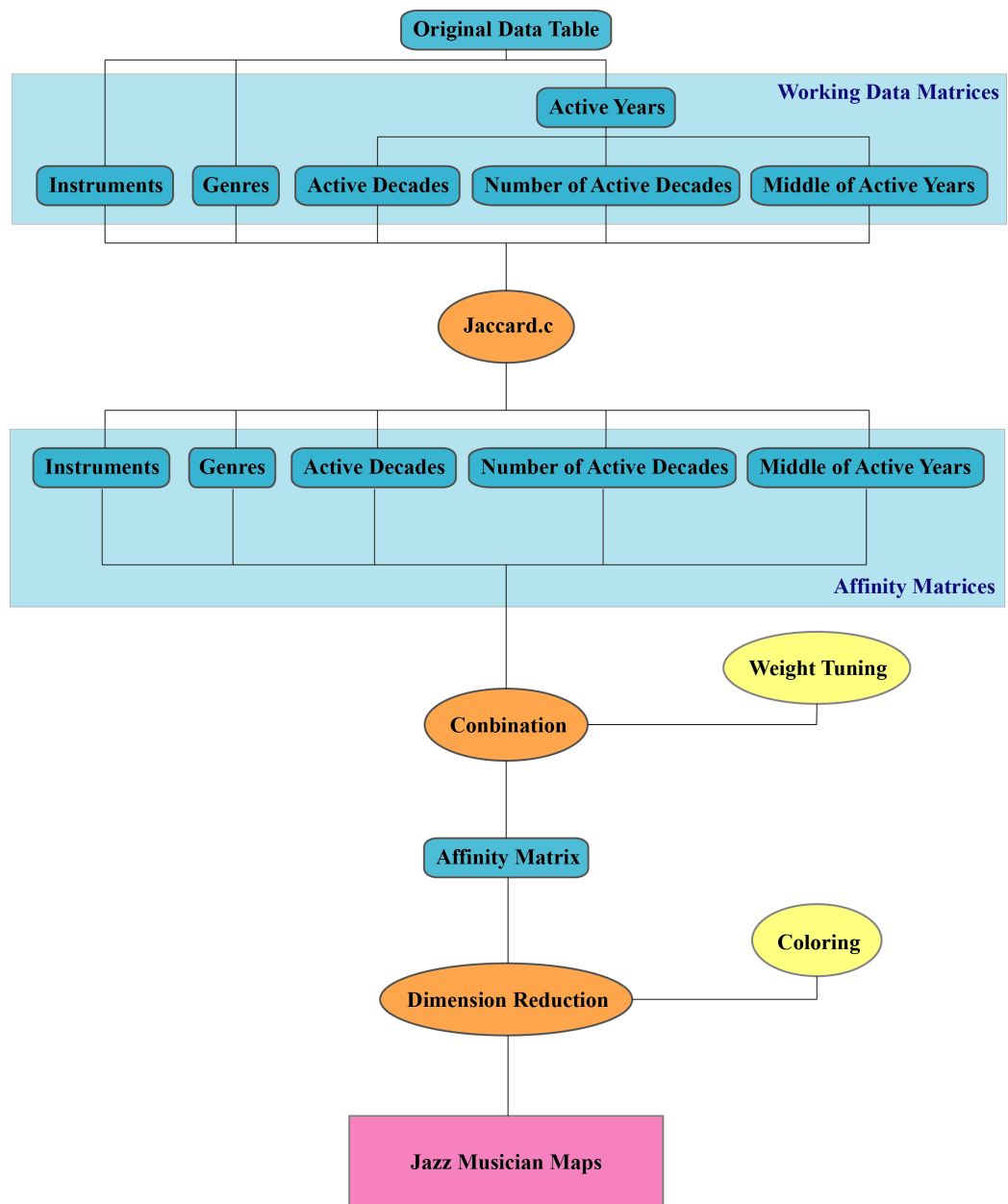
Figure 3.3: Data Table to Jazz Musician Maps

Chapter 4

Results and Discussion

Our first jazz musician map is Figure 4.1, and the data points on the map represent musicians. Neighboring musicians indicate the similarity of features, which consist of instruments, genres, and active years. While showing all of the names of each musician on the map is difficult and makes the map messy, we utilize *Plotly* map to view their names easily. Moreover, we create other maps that are colored by instruments, genres, number of active decades, and the middle of active years. (refer to Lu (2023) for jazz musician maps by Plotly)

4.1 Coloring by Instruments

In Figure 4.3, more cyan for percussions, more magenta for wind, and more yellow for vocal, the lower right side corner is the colors of three groups of instruments. The other colors are mixed in more than one group, and a dark color point means the musician who played the most instrument.

4.1.1 Marsalis Family

Wynton, Branford, Jason, and Delfeayo Marsalis, and their father Ellis Marsalis Jr., are members of the Marsalis family. Although Delfeayo is not included in our dataset, he is also part of the family. As shown in Table 2.1, Wynton and Branford have the same active years and play only wind instruments, but they specialize in different genres. In Figure 4.4, this similarity places them in close proximity to each other, far away from other family members. While their father Ellis Marsalis Jr. is a pianist and is colored the same as Wynton and Branford, they have far a distance because of their different instruments and active years. Jason, on the other hand, is a drummer and is therefore positioned on the right side of the map relative to his wind-instrument-playing brothers. The plot illustrates that instruments determine the position and color.

25

4.1.2 Roy Hargrove Fellows Players

Roy Hargrove and Clark Terry are trumpeters, vocalists, and flugelhorn players who promote Bebop and Hard Bop. Furthermore, they are both active from 1987 to 2015. In Figure 4.5, they are very close and almost share the same color. Moreover, the two players of similar instruments have a near distance. For example, Dizzy Gillespie is a trumpeter, vocalist, and pianist, and the difference between him and Roy Hargrove is that Roy Hargrove can play Flugelhorn. Therefore, even if they have huge differences between genres and active years, they still have a near distance in the same quadrant.

4.2 Coloring by Genres

Because we use the same processing method of instruments on genres. Thus Figure 4.6 has the same instructions as Figure 4.3, and the colors with three groups of genres are on the upper left of the map. As shown in this map, the points in white are those missing values, that is, these musicians do not have genre value.

4.2.1 Bebop and Hard Bop

Art Blakey and J. J. Johnson are both active from the 1940s to the 1990s and promote bebop and hard bop. But Art Blakey plays drums and percussions, and J. J. Johnson is a Trombonist. In Figure 4.8, they are not only in the same color but also on the top of the map in the first quadrant. The genres and active years make them a near distance. Otherwise, the instrument distances them.

4.2.2 Miles Davis

Miles Davis is an influential musician from the 1950s to the 1980s, and he also promotes many genres including Modal jazz, cool jazz, jazz fusion, and hard bop. As we know hard bop, Clark Terry is one of the most representative musicians of it, which is why they have a near distance though they are not in the same color but tone. In Figure 4.8, they are on the top of the map as PC2-axis over 0.1 area.

4.3 Active years

Active years are a crucial piece of information that we use different approaches to handle it. In Section 3.1.2, we create the number of decades and the midpoint of active years as new features, both of which are single values, making them suitable for coloring data points. In Figure 4.11, the dark colors represent earlier musicians. In Figure 4.12, the dark colors of musicians mean they have a longer active career.

Explained Variation of PCA

On the other hand, we use both PCA and t-SNE to visualize. While most t-SNE maps display more clearly groups, PCA maps can explain the proportion of variance from high-dimension. (refer to Table 4.1) We decide to use t-SNE maps as the main plots. However, Figure 4.6 is a special case in which PCA provides a better graphic and even explains more than 50% of variation by PC1 and PC2. Therefore, PCA and t-SNE are effective visualization methods for our data.

Table 4.1: Proportion of Variance

	PC1	PC2	Cumulative first two PCs (PC1+PC2)
Instruments	0.2254	0.1528	0.3782
Genres	0.3817	0.1606	0.5423
Number of active decades	0.2336	0.1694	0.4029
Middle of active years	0.2423	0.1847	0.4270

Comparison

To compare with *Linked Jazz*, we take our Figure 4.3 and Figure 4.13. In Figure 4.13, this network graph displays the relationship between each pair of musicians by a line, but our map illustrates the similarities between all of them by the distances. In addition, we color the musicians by instruments, which makes the map clearly display the musicians who play similar instruments.

Remark

Indeed, our visualizations prioritize highlighting a primary feature while also incorporating other aspects simultaneously. The objective of these visualizations is not classification but rather to provide a comprehensive representation. As a result, the maps may exhibit scattered clusters and mixtures, such as intersections between clusters, and their effect is subjective. For instance, Figure 4.3 emphasizes the prioritized feature of instruments, but we can still observe the similarity in genres and active years. However, if the disparities among other features are more significant than those among instruments, the data points will be located further apart. For example, pianists Oliver Jones and Ellis Marsalis Jr., despite sharing the same instrument, may be located at a considerable distance due to their contrasting genres and active years, potentially even belonging to different regions of the visualization.
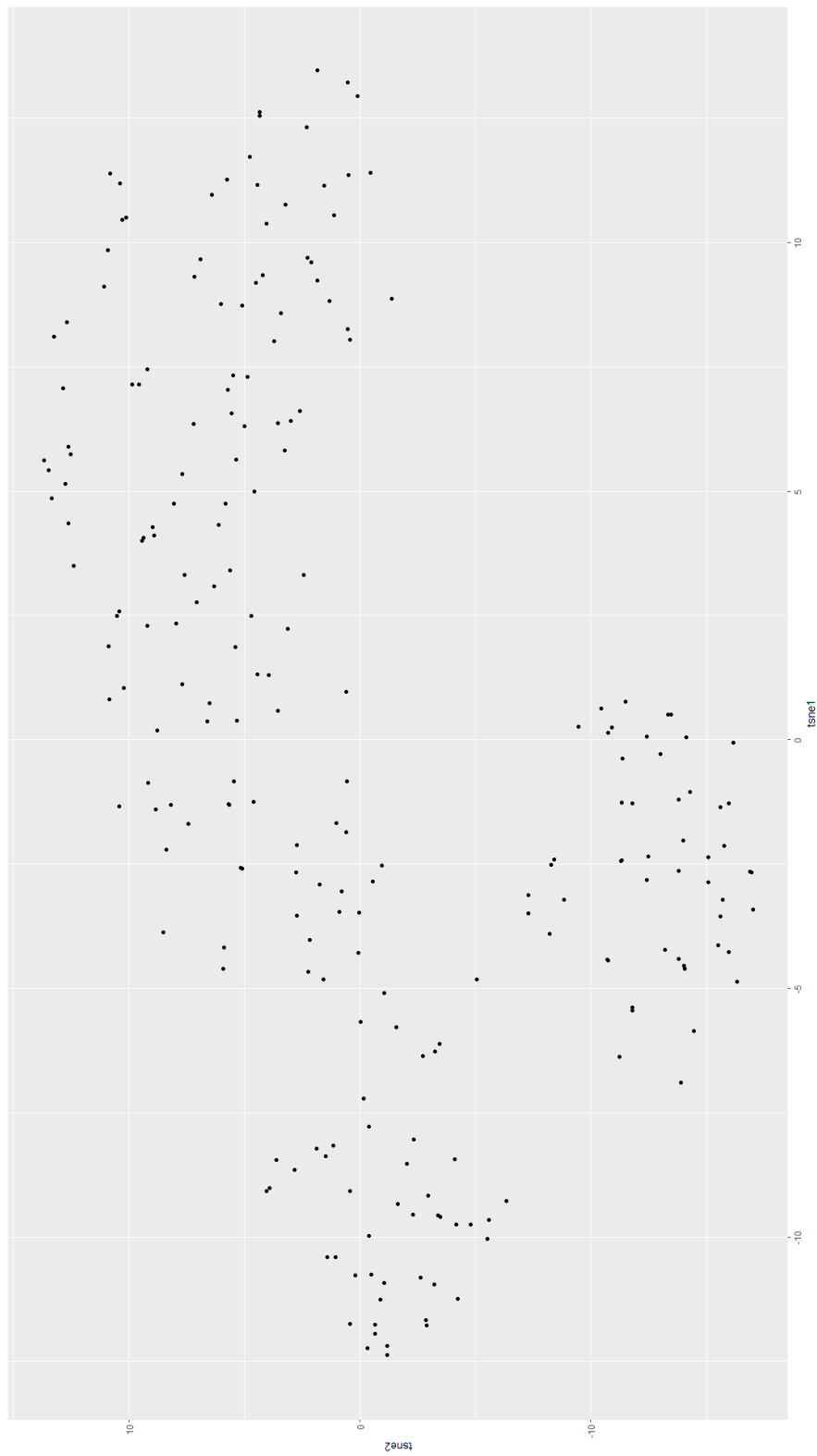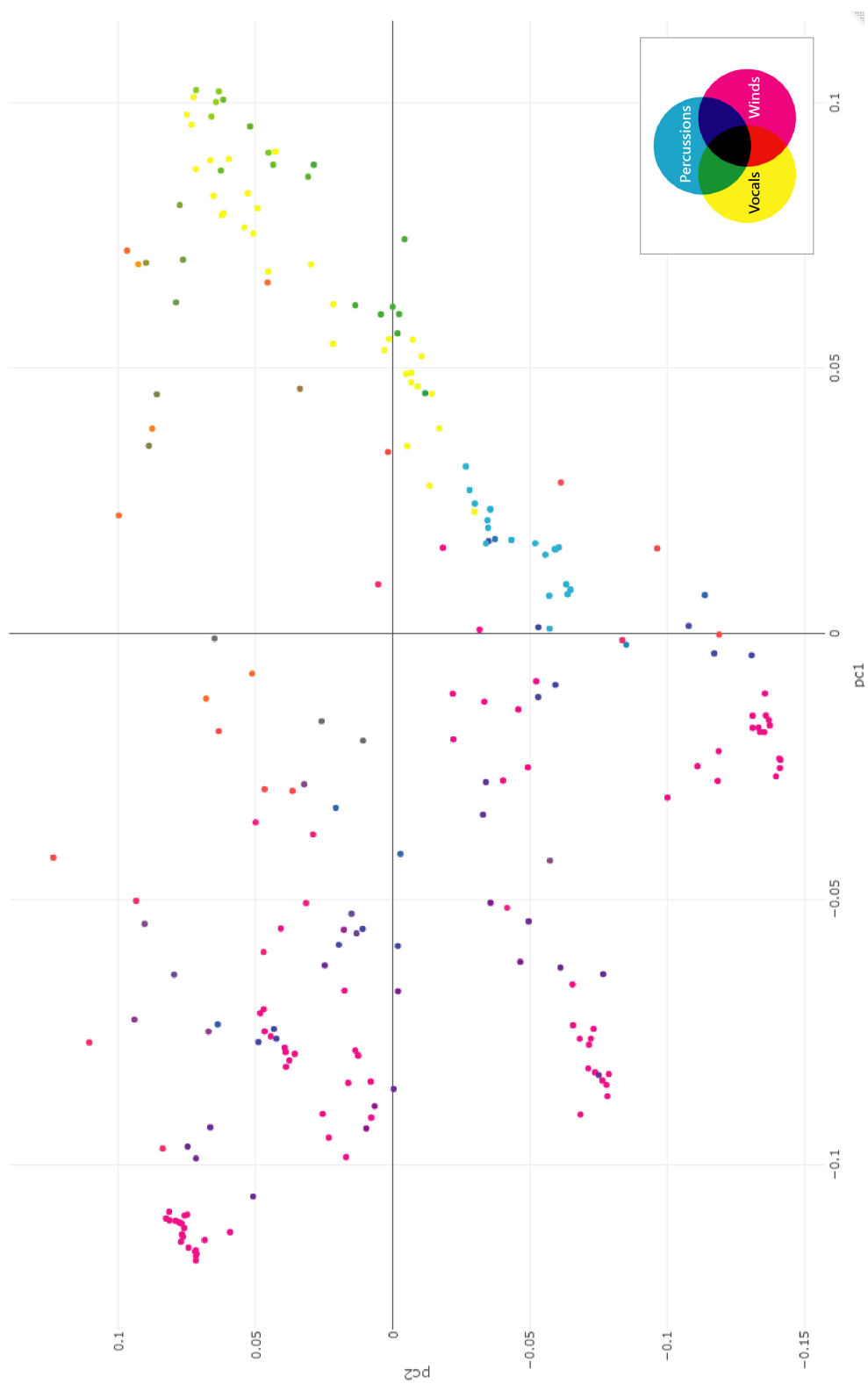
Figure 4.1: Jazz Musician Map
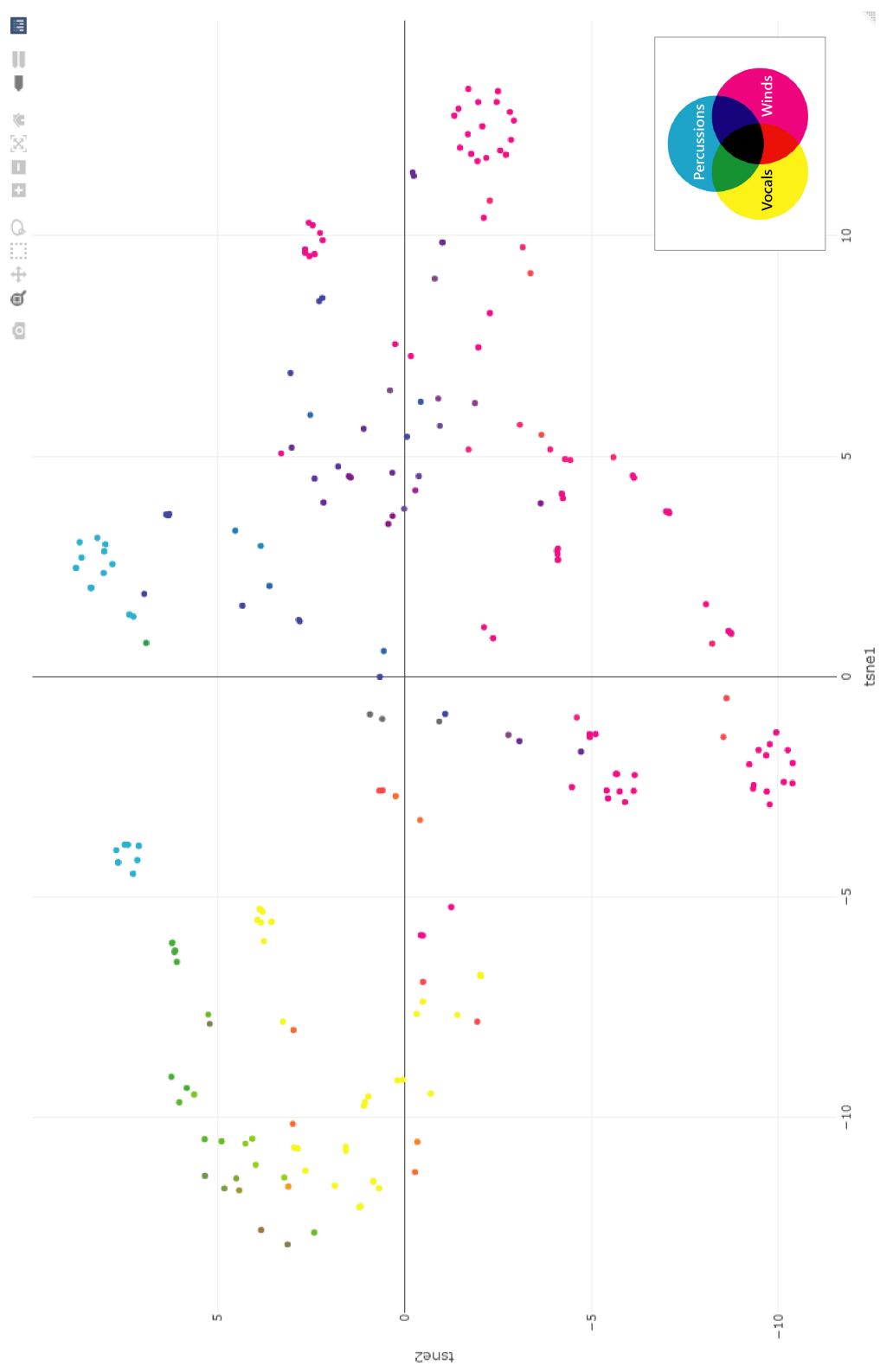
Figure 4.2: *Jazz Musician Map (Instrument, PCA)*

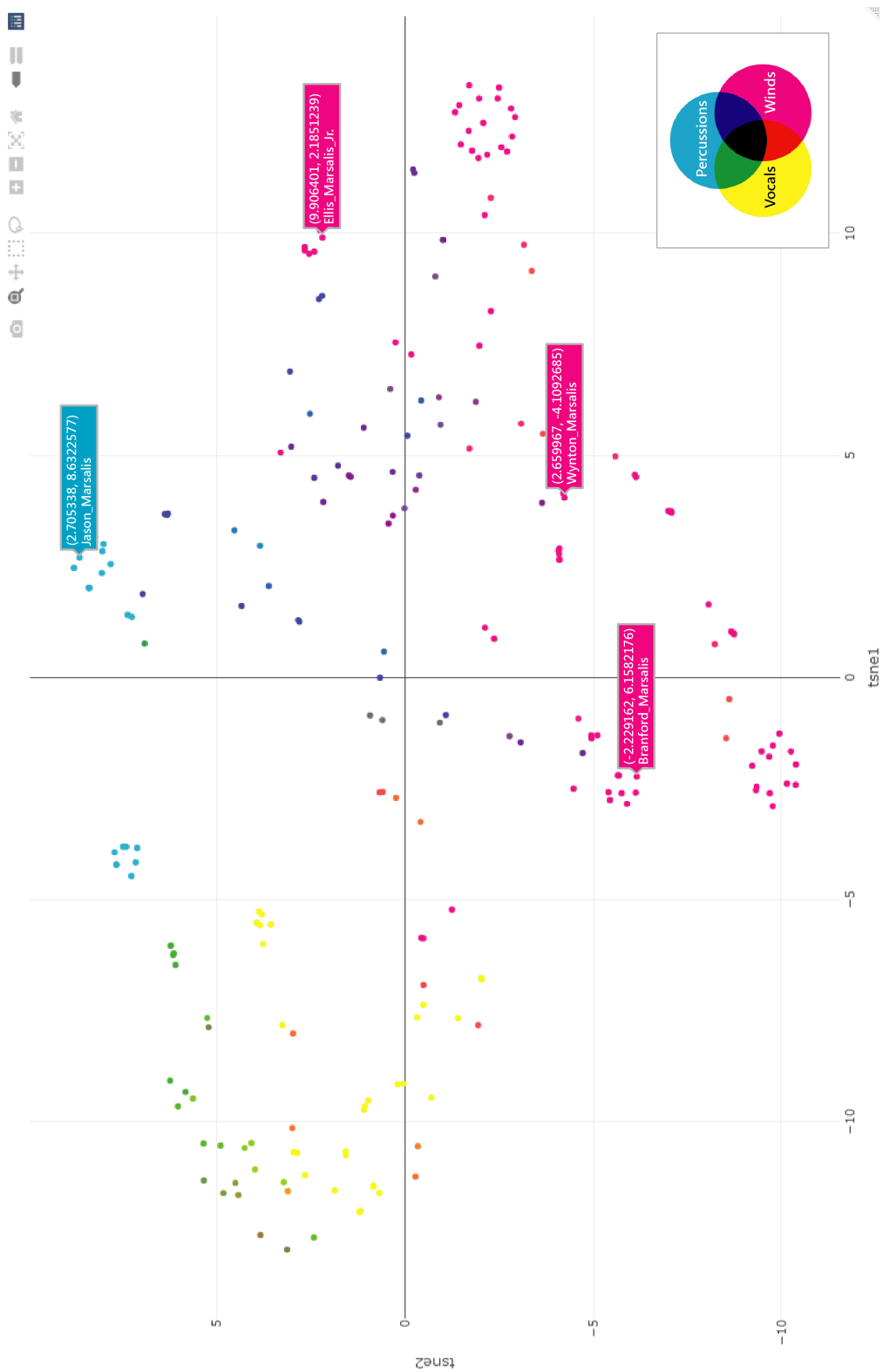**Figure 4.3:** Jazz Musician Map (Instruments, t-SNE)
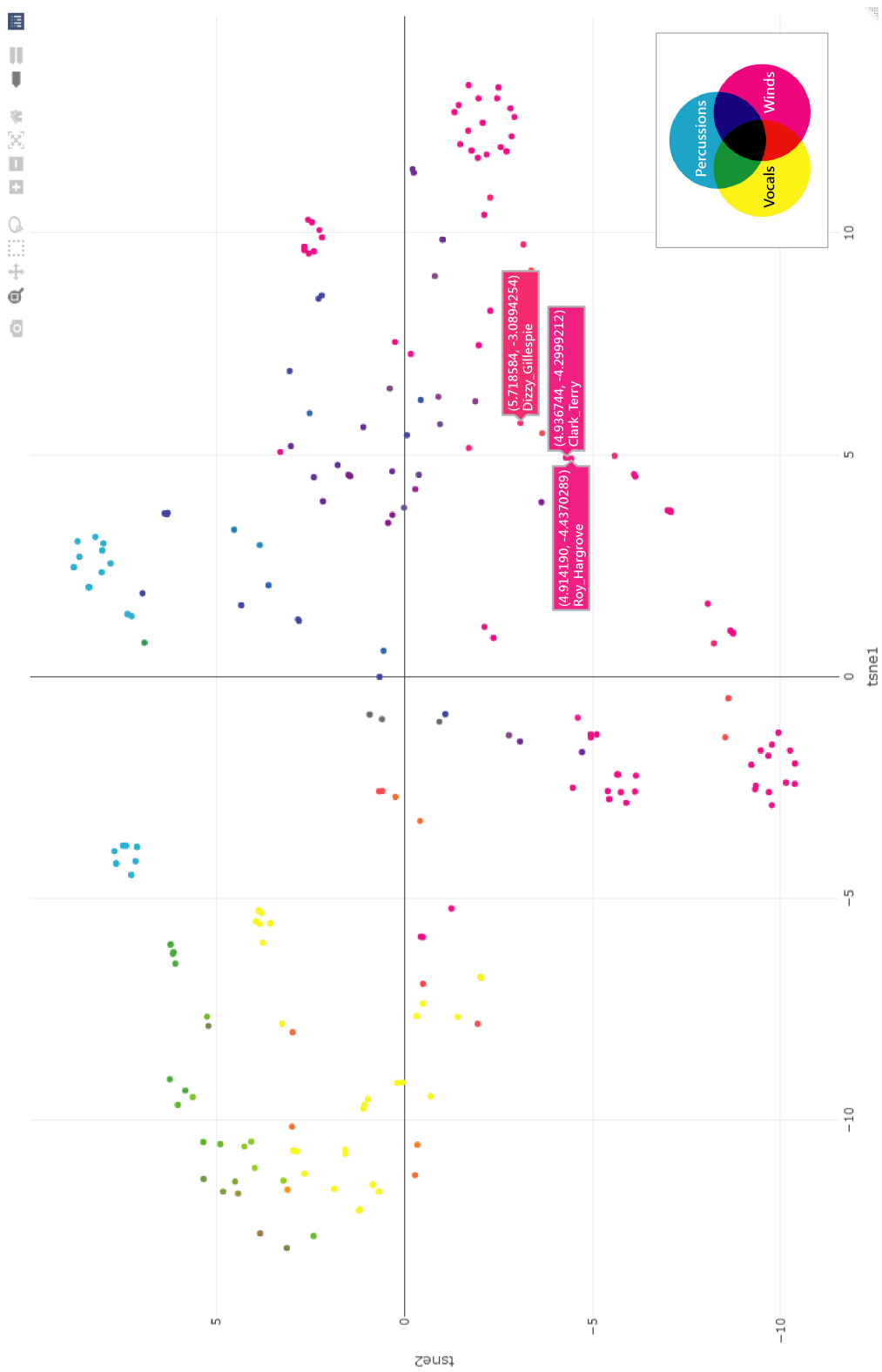
Figure 4.4: Marsalis Family Members

31

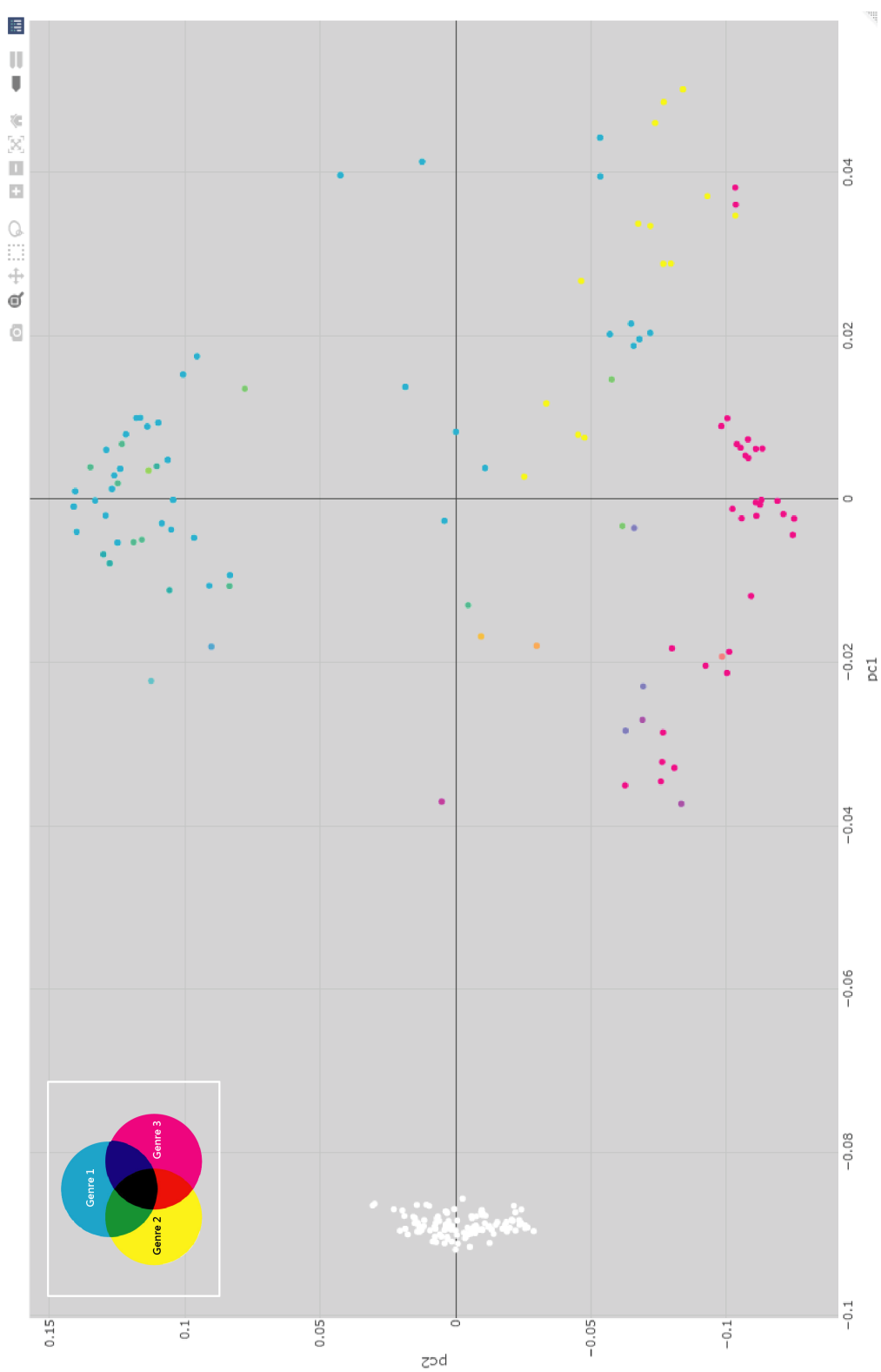**Figure 4.5:** Roy Hargrove, Clark Terry, and Dizzy Gillespie

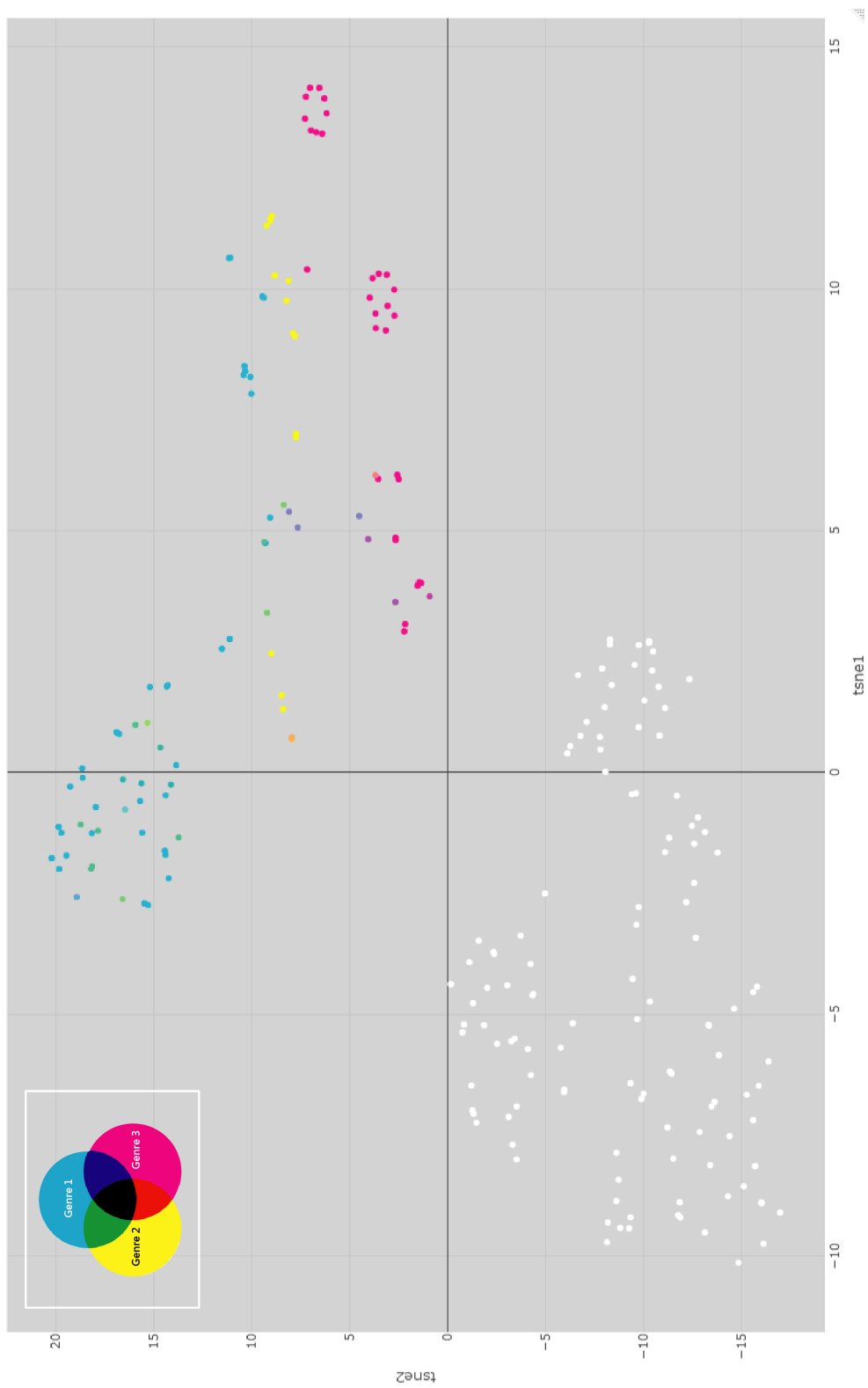**Figure 4.6:** Jazz Musician Map (Genres, PCA)

Figure 4.7: Jazz Musician Map (Genresm, t-SNE)

Figure 4.8: Hard Bop Promoters

The labelled points read:

(-0.007877695, 0.127772341)
Miles_Davis

(0.002889049, 0.126106456)
J_J_Johnson
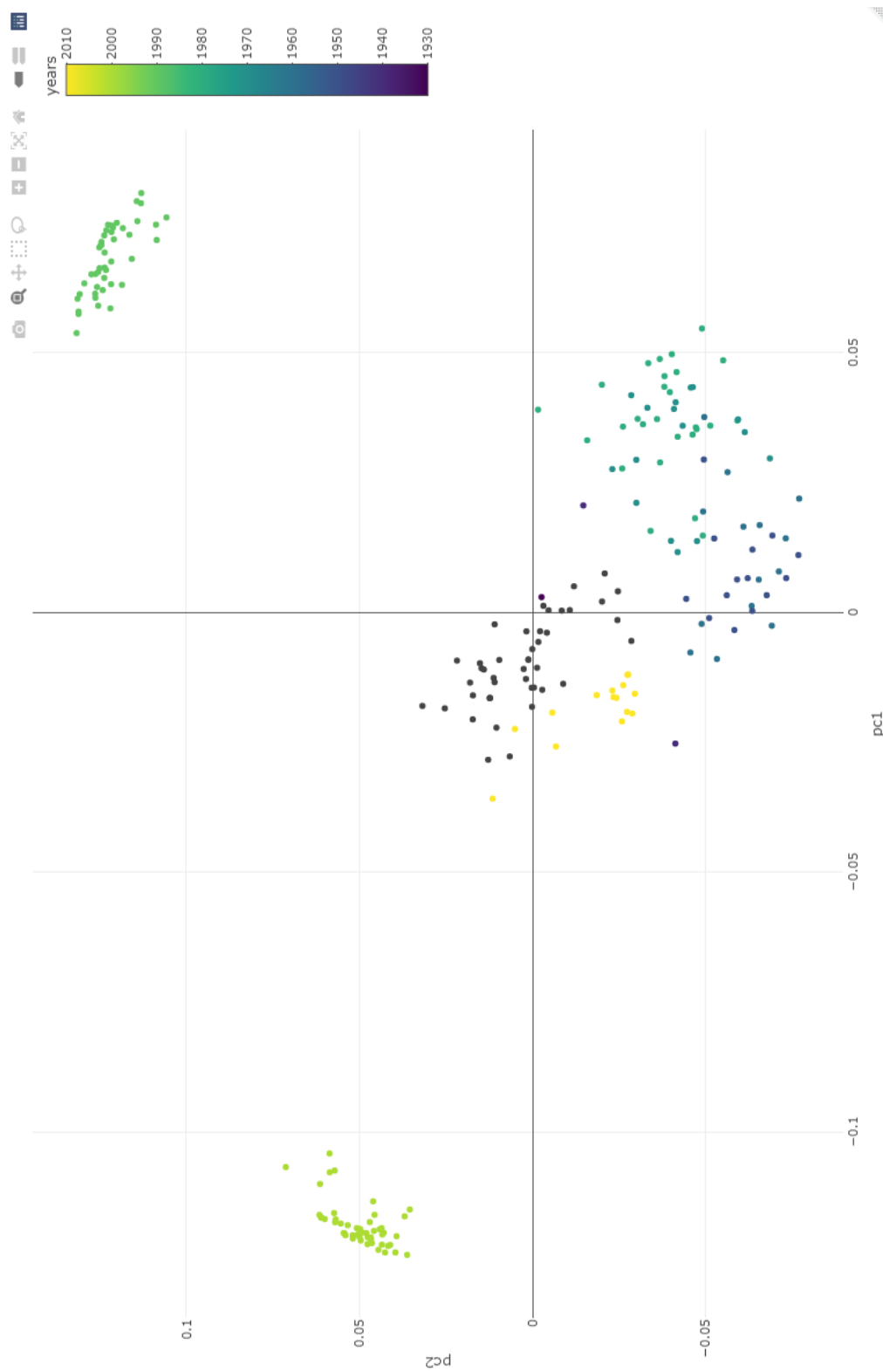
(0.009357898, 0.109809072)
Art_Blakey

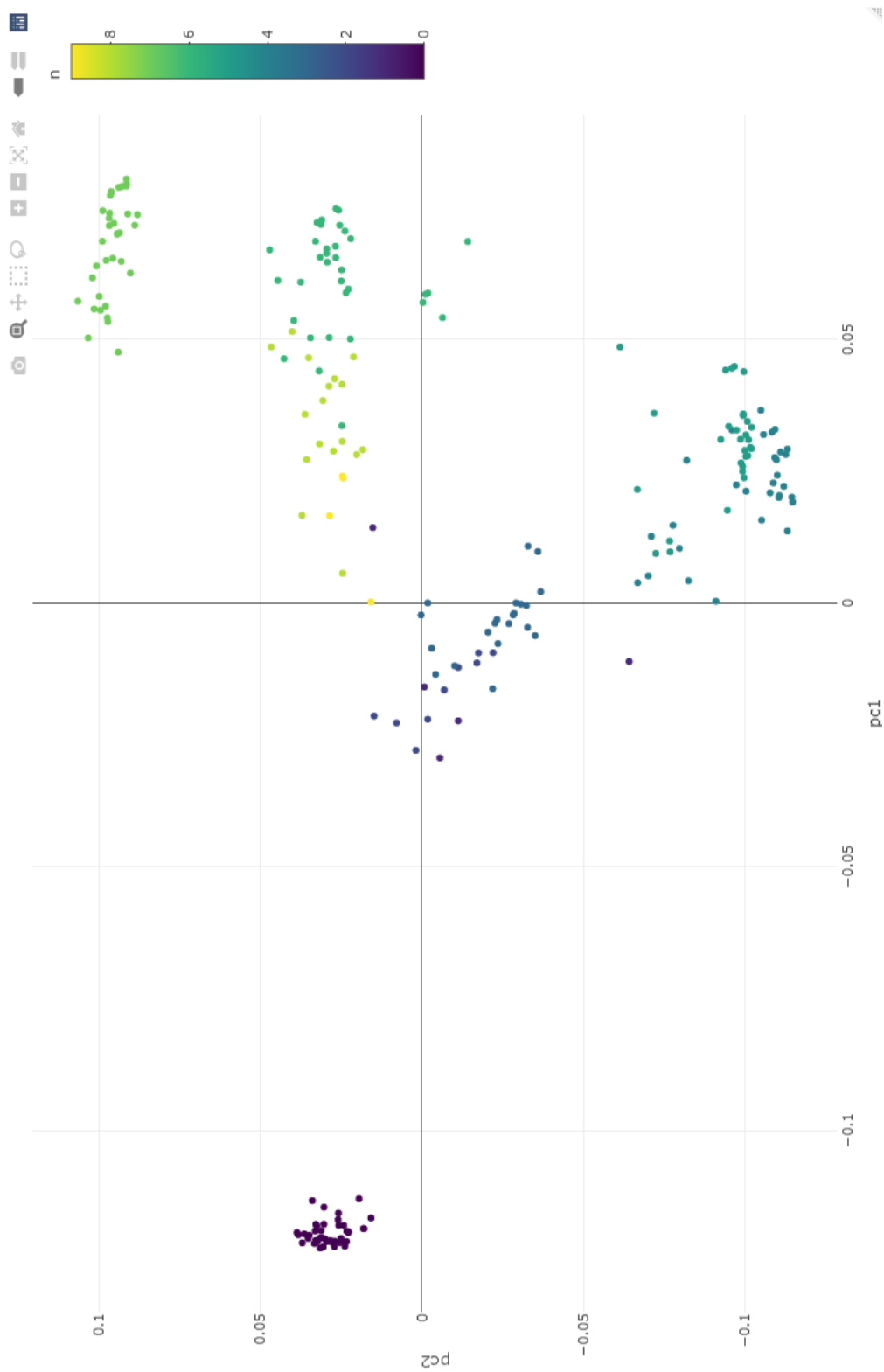**Figure 4.9:** Jazz Musician Map (Middle of Active Years, PCA)

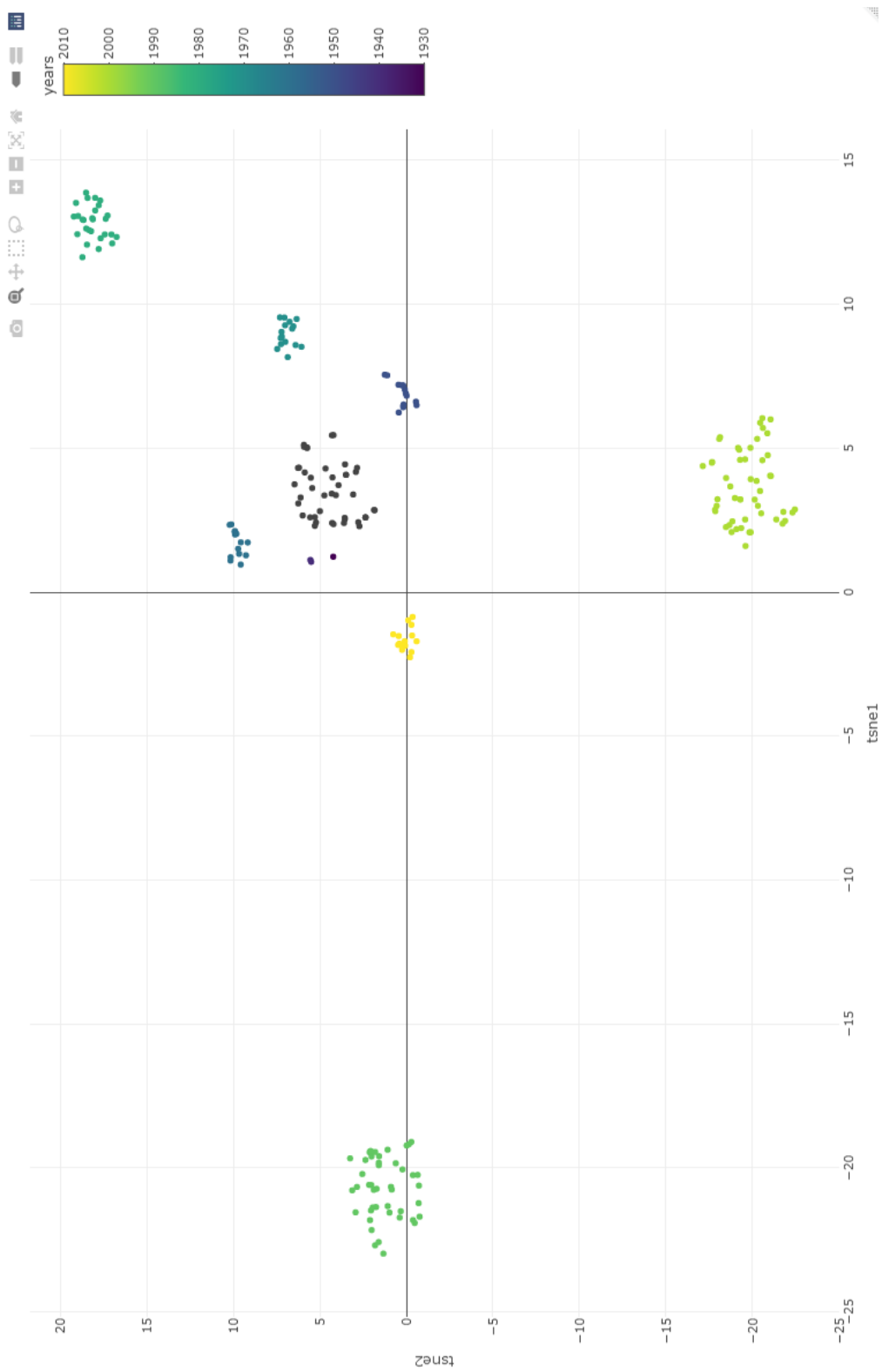**Figure 4.10:** Jazz Musician Map (Number of Active Decades, PCA)

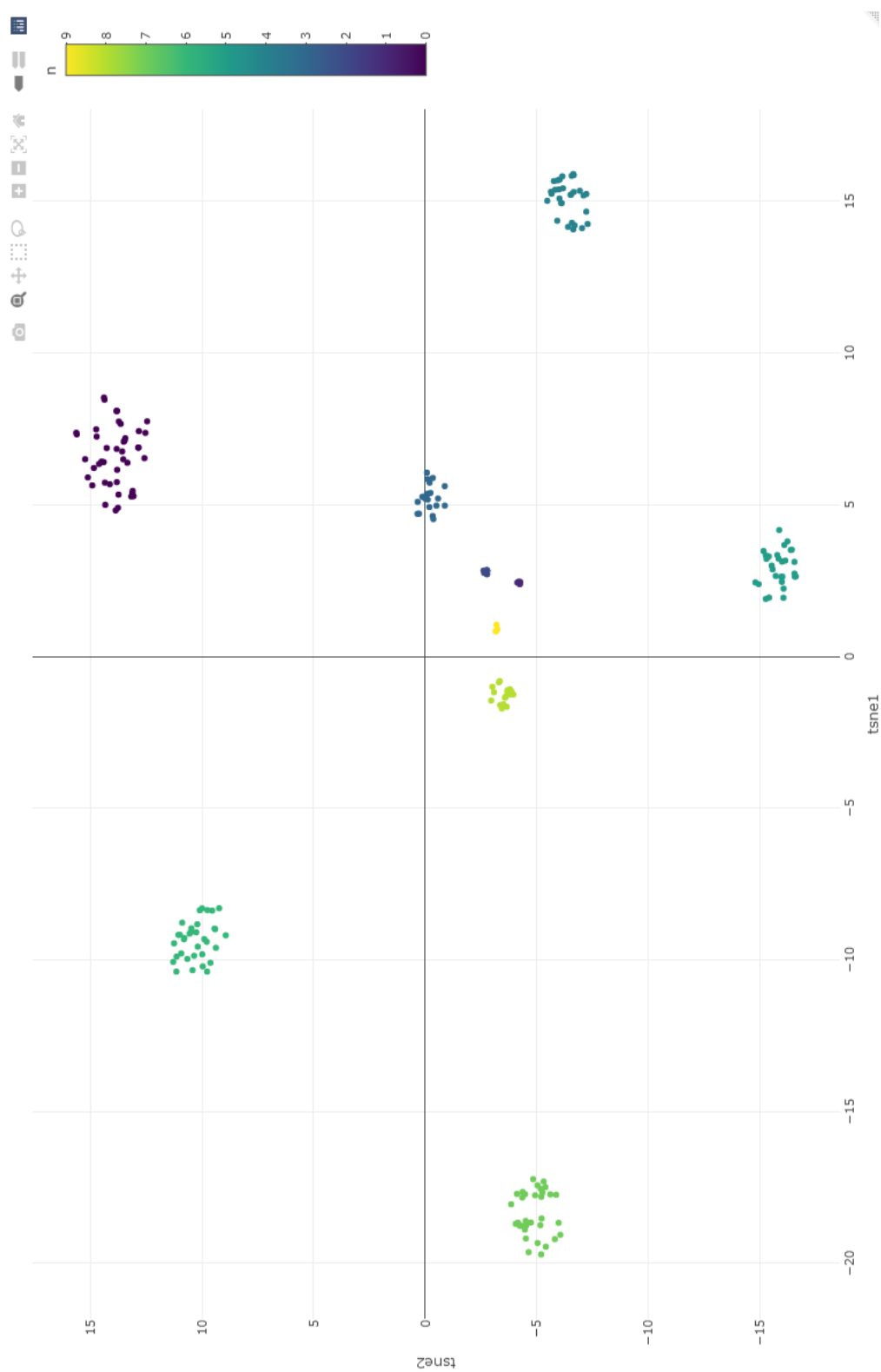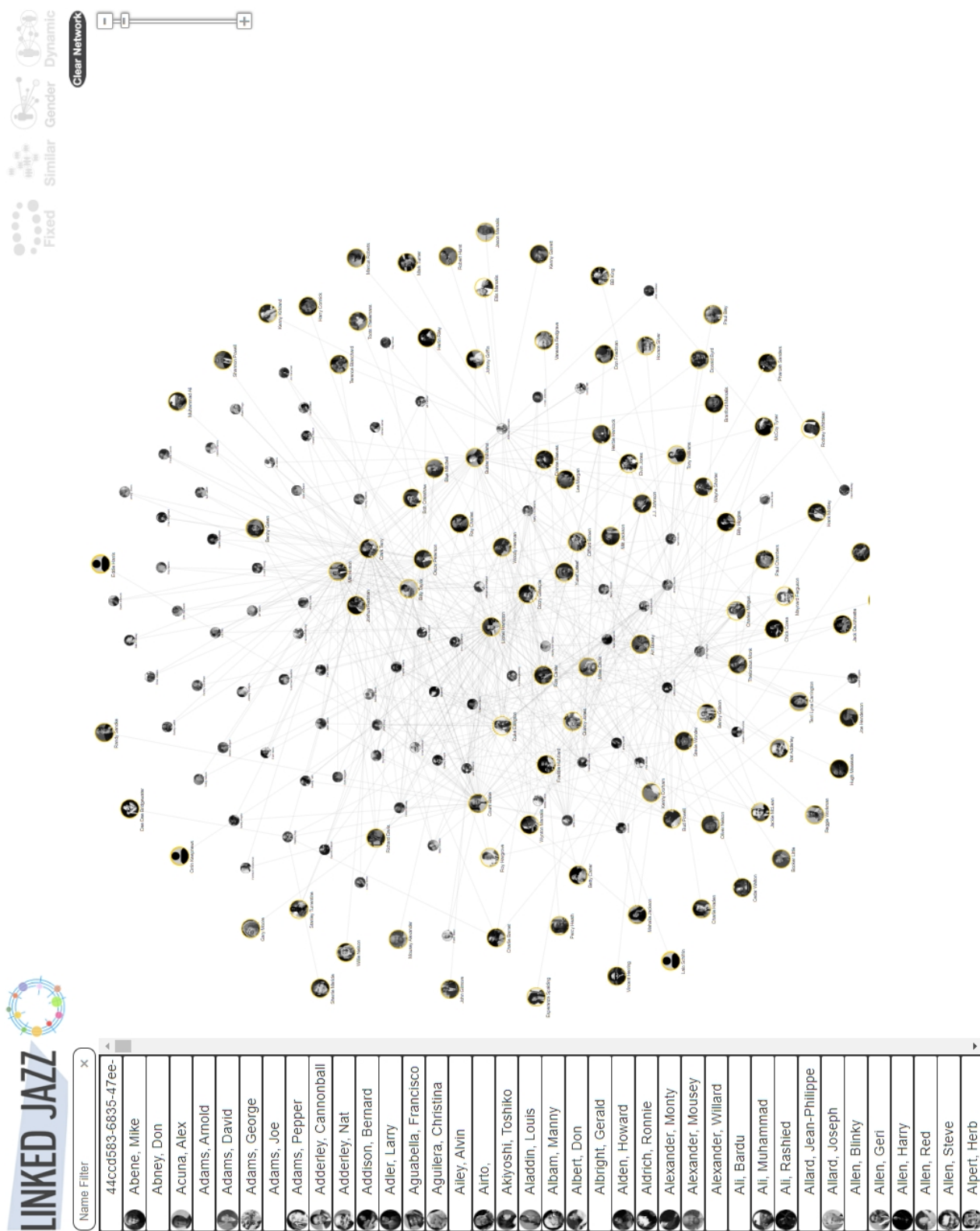**Figure 4.11:** Jazz Musician Map (Middle of Active Years, t-SNE)

Figure 4.12: Jazz Musician Map (Number of Active Decades, t-SNE)

LINKED JAZZ

Fixed · Similar · Gender · Dynamic

Clear Network

Name Filter ✕

44ccd583-6835-47ee-
Abene, Mike
Abney, Don
Acuna, Alex
Adams, Arnold
Adams, David
Adams, George
Adams, Joe
Adams, Pepper
Adderley, Cannonball
Adderley, Nat
Addison, Bernard
Adler, Larry
Aguabella, Francisco
Aguilera, Christina
Ailey, Alvin
Airto,
Akiyoshi, Toshiko
Aladdin, Louis
Albam, Manny
Albert, Don
Albright, Gerald
Alden, Howard
Aldrich, Ronnie
Alexander, Monty
Alexander, Mousey
Alexander, Willard
Ali, Bardu
Ali, Muhammad
Ali, Rashied
Allard, Jean-Philippe
Allard, Joseph
Allen, Blinky
Allen, Geri
Allen, Harry
Allen, Red
Allen, Steve
Alpert, Herb

Figure 4.13: Linked Jazz Network Graph (our musicians list)

Chapter 5

Conclusions

The thesis of our study is to construct jazz musician maps. We begin with an overview in Figure 4.1, and then we color and assign weights to instruments, genres, the number of active decades, and the middle of active years to create Figure 4.2-4.12. These maps utilize distance to illustrate the similarities between musicians. Specifically, we place a higher weight on instruments when combining affinity matrices and reducing them into two dimensions. We then map these two-dimensional data points to represent musicians on the map, with colors corresponding to their respective instruments. As a result, musicians who play similar instruments are positioned closer to each other on the map.

When coloring the maps based on instruments, we employ three primary colors (cyan, magenta, and yellow) to represent three groups of instruments: percussions, wind instruments, and vocals. In Figure 4.3, we observe that the majority of musicians on the map are represented in magenta, indicating that wind instrument players dominate our dataset. Additionally, we mark the members of the Marsalis family in Figure 4.4 to illustrate that they are not located closely on the map, as they play different instruments. Conversely, Figure 4.5 showcases musicians such as Roy Hargrove, Clark Terry, and Dizzy Gillespie, who almost exclusively play the same instruments and are positioned very closely, even appearing on the same point.

For coloring the maps based on genres, we utilize the same three primary colors to represent three groups of genres. While we use a similar method of progression for coloring instruments and genres, it is worth noting that genres cannot be strictly classified based on themes or topics. We assign genre 1, genre 2, and genre 3 to represent these three groups. In Figure 4.8, musicians who promote the Hard Bop genre are marked with the same tone of color, and musicians who exclusively promote the same genre are located close to each other on the map.

In Figures 4.9-4.12, the maps are colored based on the number of active decades and the middle of active years. These two features are single-value variables that provide strong information in the visualization. Consequently, musicians who share the same ac-

tive year or active decade are clustered together on the maps, indicating their temporal proximity.

To compare our maps with Linked Jazz, we include Figure 4.13, which contains the same musicians as our maps and is constructed using the "Dynamic" mode of the Linked Jazz network graph. While Figure 4.13 displays pairwise relationships between musicians, our Figure 4.1 illustrates similarities in terms of instruments, genres, and active years. Our maps replace the lines in the network graph with distance, resulting in clearer and more informative visualizations.

In the case of coloring multi-value variables (instruments and genres), we introduce a novel approach that utilizes three primary colors. This method not only shows musicians who play instruments or genres from the same group but also employs a mixture of primary colors to represent musicians who can play two or three different groups. For example, musicians depicted in blue are a combination of cyan and magenta, indicating that they play both percussions and wind instruments. Similarly, the green tones represent musicians who are both vocalists and percussionists. Although there are a few instruments that do not fit strictly into these three groups, such as guitarists and producers, they are shown as belonging to the same category as vocalists and singers in our dataset. Likewise, we classify pianists as wind instrument players because most saxophonists in our dataset can also play the piano. While the classification of these instruments may appear challenging, it is based on the characteristics of our dataset and provides a reasonable grouping.

In summary, our study focuses on the creation of jazz musician maps that display relationships and similarities between musicians, with colorings based on instruments, genres, the number of active years, and the middle of active years. Our map conveys more underlying relations among musicians and maintains simplicity and clarity. We also introduce a novel method using three primary colors to represent multi-value variables.

References

[1] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific Reports*, vol. 1, no. 1, pp. 1–7, 2011.

[2] M. Aparicio and C. J. Costa, "Data visualization," *Communication Design Quarterly Review*, vol. 3, no. 1, pp. 7–11, 2015.

[3] A. I. Bacallado. "How visualization can change your life," Youtube. (Sep. 14, 2020), [Online]. Available: https://www.youtube.com/watch?v=E8k6LcYO_Vg&ab_channel=TEDxTalks.

[4] C.-H. Chen, W. K. Härdle, and A. Unwin, *Handbook of Data Visualization*. Springer Science & Business Media, 2007.

[5] *DBpedia Page for Wynton Marsalis*, Accessed: 2023-03-28. [Online]. Available: https://dbpedia.org/page/Wynton_Marsalis.

[6] A. Ferrari and M. Russo, *Introducing Microsoft Power BI*. Microsoft Press, 2016.

[7] A. Goldsher, *Hard Bop Academy: The Sidemen of Art Blakey and the Jazz Messengers*. Hal Leonard Corporation, 2002.

[8] W. Härdle, L. Simar, *et al.*, *Applied multivariate statistical analysis*. Springer, 2007, vol. 22007.

[9] *Linked Jazz*, Accessed: 2023-03-28. [Online]. Available: https://linkedjazz.org/network/.

[10] Y.-H. Lu. "Python Crawler." (2022), [Online]. Available: https://github.com/YiHsinLu/jazz_visual/blob/main/musician_data.ipynb.

[11] Y.-H. Lu. "Jazz Musician Maps (Plotly)." (2023), [Online]. Available: https://yihsinlu.github.io/Jazz.io/JazzMusicianMaps/JazzMusicianMaps.html.

[12] *Music Brainz*, Accessed: 2023-03-28. [Online]. Available: https://musicbrainz.org/.

[13] *Music Map*, Accessed: 2023-03-28. [Online]. Available: https://www.music-map.com/.

[14] *Plotly*, Accessed: 2023-05-10. [Online]. Available: https://plotly.com/r/.

[15] M. Sadiku, A. E. Shadare, S. M. Musa, C. M. Akujuobi, and R. Perry, "Data visualization," *International Journal of Engineering Research and Advanced Technology (IJERAT)*, vol. 2, no. 12, pp. 11–16, 2016.

[16] C. A. Tsao, Y.-H. Lu, J.-H. Lin, and C.-H. Shih, "Group Discussion in A408," 2023.

[17] A. Unwin, "Why is data visualization important? what is important in data visualization?" *Harvard Data Science Review*, vol. 2, no. 1, p. 1, 2020.

[18] L. Van Der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[19] G.-H. Wei, "Personal Suggestion for Jazz Musicians," 2022.

[20] *Wynton Marsalis Biography*, Accessed: 2023-03-28. [Online]. Available: https://www.pbs.org/kenburns/country-music/wynton-marsalis-biography%5C#:%5C~:text=He%5C%20is%5C%20the%5C%20only%5C%20artist,Center%5C%20in%5C%20New%5C%20York%5C%20City.

[21] *Wynton Marsalis Official Website*, Accessed: 2023-03-28. [Online]. Available: https://wyntonmarsalis.org/.

Appendix A

Jaccard with a constant

In Figure A.1, the right side shows the use of Jaccard with a constant, while the left side does not. In particular, the t-SNE plot displays that the points in cyan are closer compared to the original Jaccard, and the same applies to the points in yellow. This implies that the new version of Jaccard can bring the groups closer together. However, the connection between two musicians who do not play the same instruments may be significant in creating a more precise grouping on the map. In Figure A.2, the white points represent musicians that do not have genre values, and the right side map distances those points as a group.
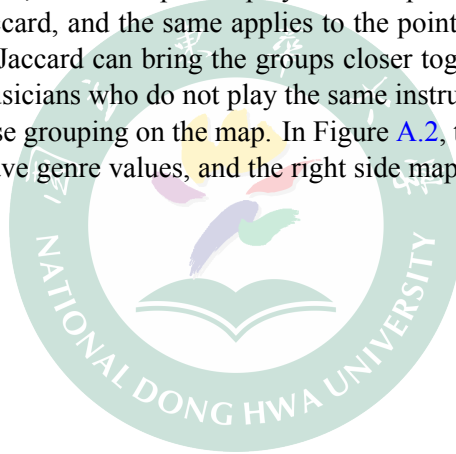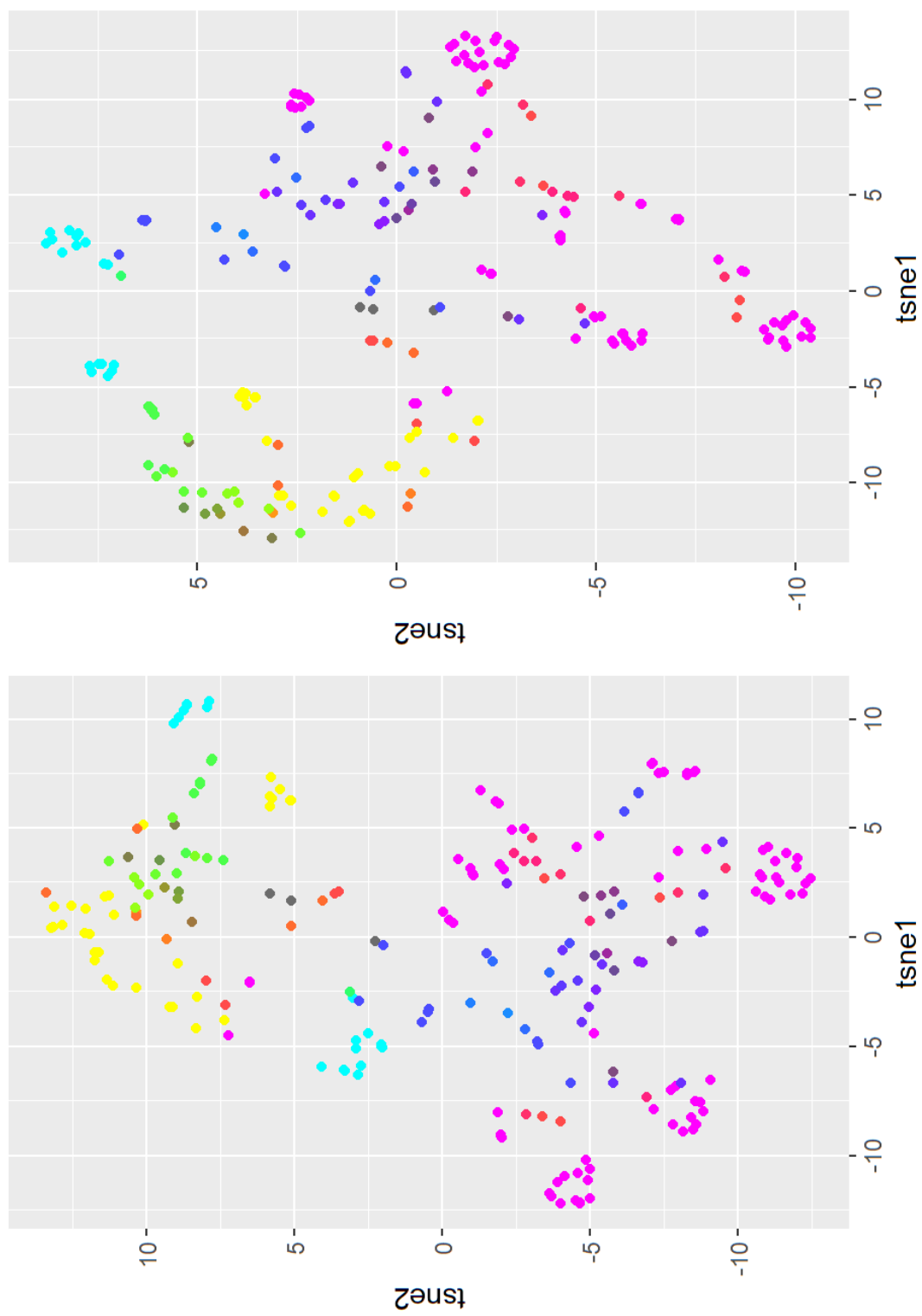
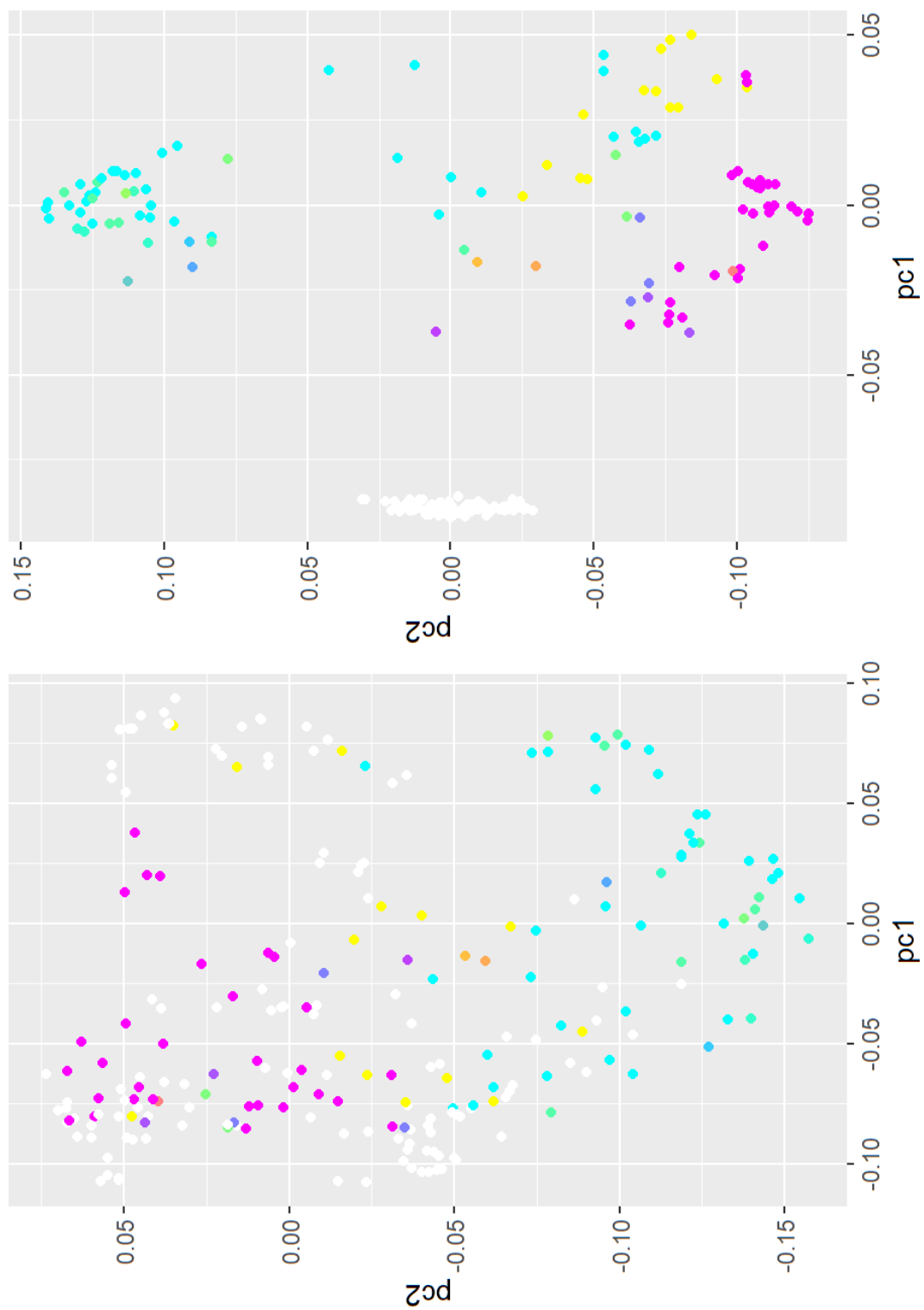**Figure A.1:** Jazz Musician Map Color by Instruments (t-SNE)

Figure A.2: *Jazz Musician Map Color by Genres (PCA)*

47