

# Autobiographical Data Collection on ■■■■

Sociology, Korea University

## 1. Variance of Interests

First, there is an argument that the research subject is interested in various fields. In this report, I analyzed categories of books the subject read and artists of his “2021 Recap” in Youtube Music. Data about books includes almost all books he has read since 01/23/2020. The data was collected by mobile application “북적북적”, which is used to write down about book the subject read. In addition, I exported it to .csv file and add column that shows category of book in KDC(Korean Decimal Classification). And I added that data using Korean National Library Open API. Youtube Music 2021 Recap is individualized music record chart in 2021. Therefore it includes his most heard songs in 2021. I collected that data using web scraping.

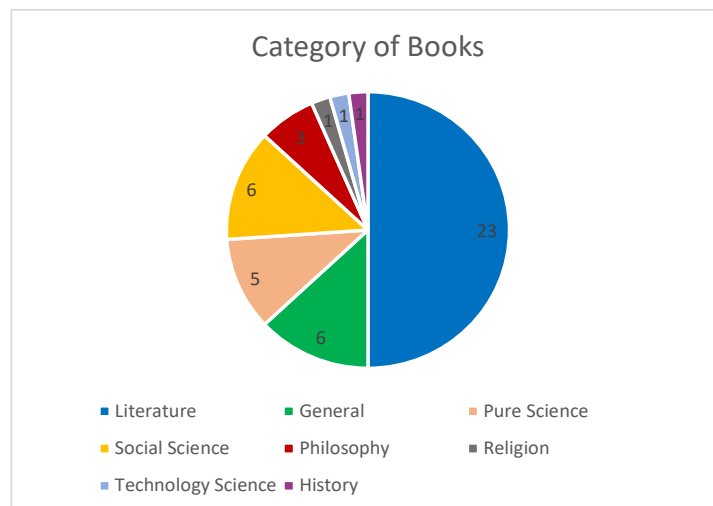


Figure 1. Categories of books subject has read

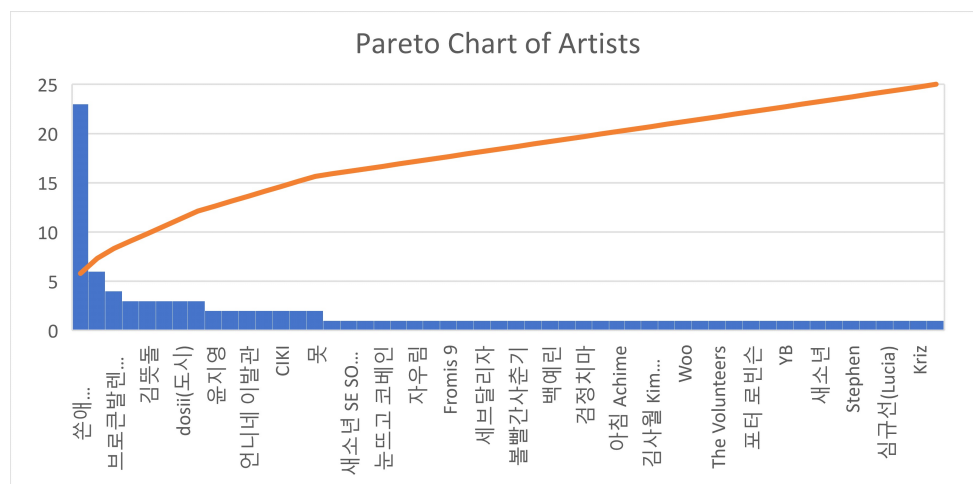


Figure 2. Pareto chart of artists who sang or created the song the subject listened. Some labels of artists are omitted for space reasons.

Categories of books(Fig 1) shows that half of books are categorized in “Literature”. At a glimpse, it seems that subject does not read various books. However, compared to recommended or real distribution of books in the library, which is proportion of Literature is highly large, variance of his book taste is normal or even large. Artists of songs are, however, truly biased. Approximately 25% of samples are created by one band, which is “THORNAPPLE(쏸애플)”. Furthermore, second place of chart is “쏸애플(THORNAPPLE)”. Yes, first place and second place is same artist. So one band accounts for 30% of samples. Therefore, the idea of his “variance of taste” is questionable.

## 2. Fields He Interested in

Second, I said that the subject is interested in current affairs, mass culture, technology, society, and computer science. This question is related to territory of his interest, while first question is about variance of it. Therefore, I used categories of books he read, which is data source used in first item. Additionally, I analyzed another data, which refers to Youtube videos he was impressed with. Categories of books data was reused, and Youtube videos data was collected using Notion, which is application used for taking notes and exporting it. As the subject has took notes about since last year, this data source includes his interest in 2021 and 2022. Then I appended categories of videos manually.

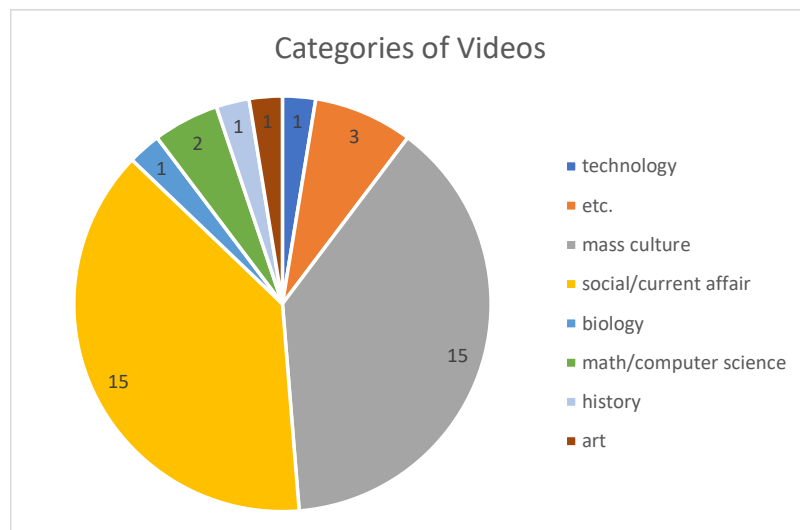


Figure 3. Categories of Videos that impressed the subject

In the data of categories of books(Fig 1), I'm not going to focus on Literature, since it has larger proportion in the population of books(e.g. recommended or real distribution of books in the library, which I mentioned above), but I note that large proportion of Literature can demonstrate interest in mass culture. Then we can see that General(G), Social Science(SS), and Pure Science(PS) occupy a large area. G includes computer science in KDC and these 6 books are all about computer science. Therefore, in this data source, G is equivalent to computer science. In



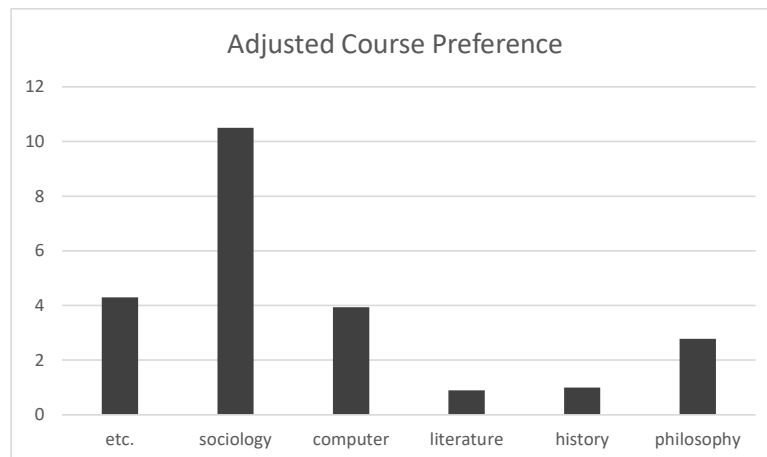


Figure 5. Sum of Credit \* Score

With word frequency data(Fig 4), we can see most frequent words like “사람(Human)”, “생각(Think)”, “우리(We)”, which is usually used. Other than that, I note that there is often-used words like “왼손잡이(Left-Handed)”, “청소년(Teenager)”, “대한민국(Republic of Korea)”, “사회(Society)”, “소아성애자(Pedophile)”, which is related to reality or society. While metaphysical words like “공포(Horror)”, “호기심(Curiosity)”, “초월(Transcendence)” are relatively less frequent. Therefore the subject may be more interested in real things than expected. And with another data source(Fig 5), sociology has high rate than other type of lectures. At a first sight, it may seem that he has a lot of concern on society so that his distance of reality is close. However, this is because he majored in sociology. Therefore, this data does not prove whether he cares reality(society in this case) or not. Thus, if we want to know his distance from reality, we should compare his data to other sociology students. Nevertheless, we can assume that the subject may be more socially involved than expected.