

# K-Pop Data Analysis

Christine P. Chai  
cpchai21@gmail.com

June 11, 2025

Starting in 2024.

## Executive Summary

Write something here

## Disclaimer

This manuscript is written solely by the author, not by ChatGPT or any other generative AI. The opinions and views expressed in this manuscript are those of the author, and do not necessarily state or reflect those of any institution or government entity.

## 1 Introduction

**Important: Write about why K-pop music is so popular across the globe.**

K-pop music (Korean popular music) has emerged popularity worldwide since the early 2010's (Khiun, 2013; Sun, 2022). K-pop's international popularity stems from its appealing music, eye-catching performances, and the strategic use of social media, fostering a global community of dedicated fans who actively promote the artists and genre (Kim and Kwon, 2022; Chen, 2023). Therefore, K-pop enjoys fame not only across East Asian countries like China and Thailand (Malik, 2023), but also in the United States<sup>1</sup> and Western Europe<sup>2</sup> (Miroudot, 2024).

The paragraph above is too short. Need to add more content.

**Then write about the author's motivation**

The author became interested in K-pop music from the debut of Tzuyu (Chou Tzu-Yu, 周子瑜).<sup>3</sup> Tzuyu is originally from Taiwan, the country in which the author grew up. In 2015, Tzuyu participated in the South Korean reality television show *SIXTEEN*,<sup>4</sup> and eventually got added to the newly-formed girl group *TWICE*.<sup>5</sup> In early 2016, Tzuyu was forced to apologize after she raised the Taiwan flag in a Korean entertainment

---

<sup>1</sup><https://bit.ly/4iIHpor>

<sup>2</sup><https://nolae.eu/blogs/news/diese-k-pop-konzerte-werden-2025-in-europa-stattfinden>

<sup>3</sup><https://en.wikipedia.org/wiki/Tzuyu>

<sup>4</sup>[https://en.wikipedia.org/wiki/Sixteen\\_\(TV\\_program\)](https://en.wikipedia.org/wiki/Sixteen_(TV_program))

<sup>5</sup><https://en.wikipedia.org/wiki/Twice>

show.<sup>6</sup> The flag controversy incident made headline news in Taiwan,<sup>7</sup> and it was estimated to bring in 500,000 votes for the 2016 Taiwan presidential election.<sup>8</sup>

## Unfinished below

Then in 2017, another Taiwanese girl, Snowbaby (蔡瑞雪),<sup>9</sup> joined the Korean live reality show *Idol School* (偶像學校).<sup>10</sup> At the end, the show would select nine winners to form the girl group *fromis\_9*.<sup>11</sup> This also generated lots of discussion in the Mandarin-speaking community.<sup>12</sup> Although Snowbaby was eliminated in the middle of the show, she still deserves praise for her courage to participate as an international contestant.<sup>13</sup> Since Snowbaby graduated from the same high school as the author did,<sup>14</sup> Snowbaby's experience motivated the author to learn more about K-pop.

Survivorship bias is prevalent in the entertainment industry, and the K-pop genre is no exception (Lockwood, 2021). To become a K-pop idol in Korea, aspiring kids usually start as a trainee at an entertainment company in their early teens (Lee and Jin, 2019). The trainees take vocal and dance lessons, and the training process is intense and extremely competitive (Kang, 2017; Lee, 2024). (No wonder the K-pop artist performances are beautiful and well-rehearsed (Kim et al., 2021).) Very few trainees can eventually get selected to debut and perform on stage as the company's official group (Han and Pothong, 2021), and even fewer K-pop performers can achieve iconic status (Min, 2024). While a performing group enjoys the spotlight and fame, the group typically lasts only a few years before the contract ends (Cho et al., 2023).

K-pop overcrowded market (Liu, 2025).

(a lot more content here)

Need a transition paragraph to explain what we are doing in this analysis

Important: Write about the K-pop scandal revealed in 2019 and later.

In 2019, the author stopped following the shows produced by Mnet because ...

We focus on the ratings directly given by the vocal and dance instructors, rather than the published number of audience votes.

[https://en.wikipedia.org/wiki/Mnet\\_vote\\_manipulation\\_investigation](https://en.wikipedia.org/wiki/Mnet_vote_manipulation_investigation)

A K-pop vote manipulation scandal was surprisingly revealed in 2019, starting with the *Produce X 101*<sup>15</sup> and the mysterious 29,978 number.<sup>16</sup> After the producer Mnet<sup>17</sup> published the number of votes each contestant received, people calculated the difference between rankings and noticed a strange pattern of the numbers. The difference was exactly 29,978 votes for five intervals among the rankings from 1st to 10th. The pattern seemed generated by some mathematical function, and it was nearly impossible to be a coincidence. Hence people suspected that the voting results were manipulated by Mnet.<sup>18</sup>

Finally, Mnet admitted to manipulating the votes in the *Produce 101* series and the subsequent reality shows, including *Idol School*.<sup>19</sup> The show producers rigged the votes in return for financial favors, resulting in not

---

<sup>6</sup><https://bit.ly/3DOcNIP>

<sup>7</sup><https://bit.ly/4k5j7ps>

<sup>8</sup><https://bit.ly/3CUQWsK>

<sup>9</sup>Snowbaby's YouTube channel: <https://www.youtube.com/@snowbaby>

<sup>10</sup>[https://en.wikipedia.org/wiki/Idol\\_School\\_\(2017\\_TV\\_series\)](https://en.wikipedia.org/wiki/Idol_School_(2017_TV_series))

<sup>11</sup>[https://en.wikipedia.org/wiki/Fromis\\_9](https://en.wikipedia.org/wiki/Fromis_9)

<sup>12</sup><https://www.epochtimes.com/b5/17/7/2/n9346573.htm>

<sup>13</sup><https://bit.ly/41p3pym>

<sup>14</sup><https://bit.ly/424u3gv>

<sup>15</sup>[https://en.wikipedia.org/wiki/Produce\\_X\\_101](https://en.wikipedia.org/wiki/Produce_X_101)

<sup>16</sup><https://www.koreaboo.com/news/produce-x-101-rigged-votes-final-members/>

<sup>17</sup>[https://en.wikipedia.org/wiki/Mnet\\_\(TV\\_channel\)](https://en.wikipedia.org/wiki/Mnet_(TV_channel))

<sup>18</sup><https://bit.ly/4iWtPh0>

<sup>19</sup><https://www.popdaily.com.tw/korea/846603>

only unfair competition but also employment fraud (Lee and Zhang, 2021). These entertainment agency representatives were charged with bribery, fraud, and sabotage (Yoshimitsu, 2020).

*Idol School*: Vote Manipulation Investigation (2019)  
<https://www.ptt.cc/bbs/KoreaStar/M.1624467107.A.D7F.html>

What was the penalty of these entertainment agency representatives?  
<https://www.ptt.cc/bbs/KoreaStar/M.1680484737.A.28A.html>

Need academic citations, not just news links.

Impact on the K-pop industry

Impact on the trainees who lost the chance to debut

After the 2019 scandal, Mnet has been under controversy but is still actively producing K-pop dance survival shows.<sup>20</sup> Recent works of Mnet include *Kingdom: Legendary War* (2021)<sup>21</sup> and *Stage Fighter* (2024).<sup>22</sup>

## 1.1 Technical Narrative

This manuscript is created using R Markdown (Allaire et al., 2024)<sup>23</sup> for reproducible data analysis, just like our earlier technical report about the education in Taiwan (Chai, 2024). We have posted our code and data on GitHub,<sup>24</sup> so readers can download the GitHub repository and play with the script themselves.

The rest of this manuscript is organized as follows.

e.g. Chapter 23 does something.

## 2 *Idol School* Dataset (2017)

*Idol School* (偶像學校) (2017)

Emphasize that *Idol School* did not require vocal or dance experience and was willing to train the participants from scratch. Despite the low barrier to entry, many participants in the reality show had previously trained under various entertainment companies.<sup>25</sup>

In the live reality show *Idol School*, nine winners were selected to form the girl group *fromis\_9*.<sup>26</sup> This girl group debuted in 2018 and remained active until the contract with Pledis Entertainment ended in 2024. In January 2025, five members of the group signed a new contract with ASND.<sup>27</sup>

Need to write the data description

Wikipedia data: [https://en.wikipedia.org/wiki/List\\_of\\_Idol\\_School\\_contestants](https://en.wikipedia.org/wiki/List_of_Idol_School_contestants)

### 2.1 Read in the *Idol School* Dataset

We manually copy-pasted the contestant data from Wikipedia into a Microsoft Excel workbook (.xlsx), and used the R package `readxl` (Wickham and Bryan, 2023) to load the dataset. A main advantage of .xlsx over .csv is that we can have multiple data sheets in the same Excel file for consolidation. Moreover, Excel supports Chinese characters, so we can also include the Chinese names of each contestant. Since the English

---

<sup>20</sup><https://www.ptt.cc/bbs/KoreaStar/M.1618588754.A.7C9.html>

<sup>21</sup>[https://en.wikipedia.org/wiki/Kingdom:\\_Legendary\\_War](https://en.wikipedia.org/wiki/Kingdom:_Legendary_War)

<sup>22</sup>[https://en.wikipedia.org/wiki/Stage\\_Fighter](https://en.wikipedia.org/wiki/Stage_Fighter)

<sup>23</sup><https://rmarkdown.rstudio.com/>

<sup>24</sup><https://github.com/star1327p/K-Pop-Dataset>

<sup>25</sup>[https://kpop.fandom.com/wiki/Idol\\_School](https://kpop.fandom.com/wiki/Idol_School)

<sup>26</sup>[https://en.wikipedia.org/wiki/Fromis\\_9](https://en.wikipedia.org/wiki/Fromis_9)

<sup>27</sup><https://kpop.fandom.com/wiki/ASND>

translation of Korean names look similar to each other (Kim, 2020), we also include the date of birth (DOB) to make it easier to uniquely identify each contestant. For those who are able to read Chinese, we put each contestant's name in Chinese characters as well.

Specify the column names we included, also the column names we printed here.

Add the metadata in the Excel file or the Appendix ?!

Currently I prefer adding the metadata in the Excel file for proximity to the data itself.

Show the first 10 records as a snapshot of the dataset.

```
library(readxl)
idol_school = read_excel("UNFINISHED_Idol_School_Dataset.xlsx",
                        sheet="Idol_School_Dataset")

# Date of birth (DOB) should be date only, not a full timestamp.
idol_school$DOB = as.Date(idol_school$DOB)

columns_to_show = c("Name_Chn", "Name_Eng", "DOB",
                    "Vocal", "Dance", "Physical", "Overall")

idol_school[1:10, columns_to_show]
```

```
## # A tibble: 10 x 7
##   Name_Chn Name_Eng      DOB      Vocal Dance Physical Overall
##   <chr>    <chr>    <date>    <dbl> <dbl>    <dbl>    <dbl>
## 1 NATTY    NATTY      2002-05-30  9.8   8        8.1      8.63
## 2 劉怡伶    Tasha      1993-10-11   8     9.5      8        8.5
## 3 李采映    Lee Chae Young 2000-05-14  8.5   8.5      7.5      8.17
## 4 宋河英    Song Ha Young 1997-09-29  8.6   5.9      9.8      8.1
## 5 金恩書    Kim Eun Suh   2000-11-14  6.3   6.9     10       7.73
## 6 金明智    Kim Myong Ji  1997-10-09  5.5   7.9      8.2      7.2
## 7 張圭惻    Jang Gyuri   1997-12-27  7.2   7.1      7        7.1
## 8 朴宣      Park Sun     2004-05-25  9.5   6.1      5.5      7.03
## 9 李悠汀    Lee Yoo Jeong 1997-02-26  5.8   6.2      9        7
## 10 金娜妍    Kim Na Yeon  1996-05-15  8.3   6        6.4      6.9
```

Explain why we removed the 41st contestant whose scores were all zeros.

The 41st contestant, Som Hye In (慎惠仁), left the *Idol School* show due to health reasons. She was unable to complete the basic test, so her score was zero in all three categories (vocal, dance, and physical).

```
# UNFINISHED HERE
# We MUST remove the 41st contestant's scores (all zeros)!!
idol_school = idol_school[1:40,]
```

## 2.2 Idol School: Exploratory Data Analysis

Context: Write about how the vocal, dance, and physical scores were evaluated.

Physical testing contains a group exercise and an individual exercise.

Also mention the top performers in overall scores and in each category.

Most of the top performers had experience as a trainee under an entertainment company, and some even had debuted before.

### Then why did they join the Idol School reality show?

Unfortunately, Snowbaby (蔡瑞雪) did not do well.

Snowbaby had not received any vocal or dance training in K-pop, so she was a complete beginner in the show.

### What changes did we make from the Wikipedia data?

Our presumption is that in each category, no two contestants should have the same score. However, after sorting the *Idol School* data, we found two 3.5's and two 1.2's in the physical scores. Especially that the two 3.5's belong to top-ranked contestants Bae Eun Yeong (裴恩英) and Park Ji Won (朴池原), this issue quickly caught our attention to make corrections to the data.

In the video clip, Park Ji Won (朴池原) and her partner were the first runner-up in the group physical exercise.<sup>28</sup> We are surprised that Ji Won's physical score was only 3.5. According to the video's score table for contestants ranked 11th to 20th,<sup>29</sup> Ji Won's physical score should be 6.2. The Wikipedia table shows an inconsistency in Ji Won's overall score, i.e., the average across the three categories. Ji Won's vocal score was 7.9, and her dance score was 5. These numbers seem to be reasonable for Ji Won, because she is known for excellent singing and good dancing as a performer.<sup>30</sup> Therefore, we assume both scores to be correct. If the physical score had really been 3.5, then Ji Won's overall score would be 5.47, dropping her from 13th place to the 18th. If the overall score of 6.37 had been correct, then Ji Won's physical score should be 6.2. The second scenario is more likely to be true, given the evidence we found in the video clip. Hence we corrected Ji Won's physical score to 6.2.

The two 1.2 physical scores are more difficult to check for the underlying values, probably because they occurred in two contestants of lower ranking.<sup>31</sup> The two contestants, Jessica Lee (李瑟) and Michelle White (懷特·米雪兒), ranked in the lower half of all 41 contestants in terms of the overall ability test. Both of them got eliminated in the first round, so they did not receive much attention in the show. With the help of Google Translate,<sup>32</sup> we were able to translate the image of Korean text to (readable) English. Finally, we discovered that Michelle White's physical score should be 1.3, not 1.2.

*Idol School* (2017): Videos with subtitles in Simplified Chinese are available on the Bilibili platform.<sup>33</sup>

Screenshots saved:

[https://github.com/star1327p/K-Pop-Dataset/tree/main/Idol\\_School\\_Rating\\_Screenshots](https://github.com/star1327p/K-Pop-Dataset/tree/main/Idol_School_Rating_Screenshots)

### Still need to write the description

### Consider hiding the sorted scores

```
vocal_sorted = sort(idol_school$Vocal, decreasing = TRUE)
dance_sorted = sort(idol_school$Dance, decreasing = TRUE)
physical_sorted = sort(idol_school$Physical, decreasing = TRUE)

# UNFINISHED HERE
combined_all_three = cbind(vocal_sorted, dance_sorted, physical_sorted)
sorted_scores_df = as.data.frame(combined_all_three)
```

Test

```
sorted_scores_df[1:10,]
```

<sup>28</sup>Screenshot of the group physical exercise: <https://bit.ly/4a7QT9m>

<sup>29</sup><https://bit.ly/400KUHH>

<sup>30</sup>Park Ji Won was the main vocalist in *fromis\_9*. <https://bit.ly/402yCFI>

<sup>31</sup>Physical scores of all contestants in *Idol School*: <https://bit.ly/3DRNK0Z>

<sup>32</sup><https://translate.google.com/>

<sup>33</sup><https://www.bilibili.com/video/BV1554y1C7wj/>

```
##      vocal_sorted dance_sorted physical_sorted
## 1          9.8          9.5          10.0
## 2          9.5          9.3          9.8
## 3          8.6          9.0          9.0
## 4          8.5          8.5          8.7
## 5          8.3          8.4          8.2
## 6          8.0          8.0          8.1
## 7          7.9          7.9          8.0
## 8          7.2          7.5          7.5
## 9          7.0          7.4          7.0
## 10         6.5          7.1          6.5
```

Test

```
sorted_scores_df[31:40,]
```

```
##      vocal_sorted dance_sorted physical_sorted
## 31          2.5          3.6          1.6
## 32          2.2          3.5          1.5
## 33          2.1          3.3          1.3
## 34          2.0          3.2          1.2
## 35          1.5          3.1          1.0
## 36          1.4          3.0          0.9
## 37          1.3          2.6          0.8
## 38          1.2          2.2          0.7
## 39          1.1          2.0          0.6
## 40          1.0          1.0          0.4
```

## 2.2.1 Summary Statistics

Check for the mean and median of each category score

The five-number summary refers to the five most important percentiles in the data sample – minimum, 1st quartile, median, 3rd quartile, and maximum.<sup>34</sup> The `summary` function in R outputs the five-number summary along with the arithmetic mean for the data.

```
# Combine all three summary tables
vocal_summary = summary(idol_school$Vocal)
dance_summary = summary(idol_school$Dance)
physical_summary = summary(idol_school$Physical)

score_summary = rbind(vocal_summary, dance_summary, physical_summary)
row.names(score_summary) = c("Vocal", "Dance", "Physical")
print(score_summary)
```

```
##      Min. 1st Qu. Median   Mean 3rd Qu. Max.
## Vocal   1.0   2.875   4.95 4.8850  6.425  9.8
## Dance   1.0   3.825   5.55 5.4850  7.025  9.5
## Physical 0.4   1.675   3.25 4.1925  6.425 10.0
```

<sup>34</sup>[http://en.wikipedia.org/wiki/Five-number\\_summary](http://en.wikipedia.org/wiki/Five-number_summary)

Observation:

Vocal and dance scores have 1 as the minimum score (who participated but made a blunder while performing), and the maximum score is below 10 (full mark). Physical scores have a wider range because the first place is automatically given a 10, and the other contestants' scores are calculated using the best performer as the baseline. Hence it is possible to receive a physical score below 1.

Vocal:

NATTY received 9.8 as the highest vocal score, and the instructor said her singing performance was perfect (considering that NATTY is originally from Thailand and hence not a native speaker in Korean).<sup>35</sup>

On the other hand, Lee Hae In (李海印) received a score of 1 in vocal, because she lost her voice during the singing part (which was unfortunate).

Dance:

Although the best score was only 9.5 (given to Tasha (劉怡伶)), the median score of dance is the highest among the three categories.

In K-pop, dance is an essential element for performers. (citation needed)

In the live band, everyone must dance!

The video clip showed that Lee Si An (李詩安) and Yoo Ji Na (柳知娜) struggled in the dance exercise, but it was Jung So Mi (鄭昭彌) who got the lowest score of 1.

c.f. Vocal roles are divided into main vocal, lead vocal, and sub-vocal (i.e., everyone else, or simply “vocal”).<sup>36</sup>

Generally the most challenging lines are assigned to the main vocal and the lead vocal, while the other vocals receive easier parts. (citation needed)

Physical:

Kim Eun Suh (金恩書) came in first and received the full mark of 10, while Song Ha Young (宋河英) got 9.8 as the second place.

Min = 0.4, and a total of five students had a score below 1.

Lowest median score among all three categories.

The show did not reveal how the physical scores were calculated from the raw time sustained by each participant the exercises.

### Physical: group physical exercise vs individual exercise

We suspect that the individual exercise accounted for a larger proportion of the physical score. We saw in the video that Lee Da Hee (李多熙) and Kim Na Yeon (金娜妍) came in first in the group physical exercise. However, Lee Da Hee's physical score was only 4.9, and Kim Na Yeon got 6.4 (which is good but not one of the best).<sup>37</sup>

**Remark: That's why it is important to remove the record with all zeros, otherwise the minimum score would be zero in every single category. Then we would not have noticed that the vocal and dance scores start from 1 with mere participation, but the physical scores have a different scale.**

Comparison:

The median in physical score (3.25) is lower than the median in vocal (4.95) or dance (5.55), indicating that many contestants did not do well in the physical test. In fact, some contestants had a remarkably high score in dance but a low score in physical. One example is Bae Eun Yeong (裴恩英), who scored 9.3 in dance (second place) and 3.5 in physical (slightly above the median of 3.25). Another example is Lee Hae In (李海印), who got 8.4 in dance 8.4 and only 1.8 in physical.

<sup>35</sup>NATTY's birth name is Anatchaya Suputhipong. [https://en.wikipedia.org/wiki/Natty\\_\(Thai\\_singer\)](https://en.wikipedia.org/wiki/Natty_(Thai_singer))

<sup>36</sup><https://bit.ly/3RFAHDL>

<sup>37</sup>Screenshot of the first place team in the group physical exercise: <https://bit.ly/3EDBhz5>

With a few exceptions:

e.g. Jo Yuri (曹柔理) received only 2.2 in dance but 5.9 in physical. Jo Yuri had not received any K-pop training prior to the show in 2017, but she was a long jump athlete during her school years.<sup>38</sup>

e.g. Song Ha Young (宋河英) got 5.9 in dance and 9.8 in physical. Song Ha Young was a certified aerial yoga instructor before she first appeared in the *Idol School* reality show.

Shall we create a **box plot** using `ggplot2` (Wickham, 2016) to compare the three sets of scores?

<https://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization>

### 2.2.2 Correlation between Vocal, Dance, and Physical Scores

We can obtain the pairwise correlation coefficients of each category. There is a positive association between the vocal, dance, and physical scores. We round each number to three decimal places.

```
vocal_vs_dance = round(cor(idol_school$Vocal, idol_school$Dance), 3)
dance_vs_physical = round(cor(idol_school$Dance, idol_school$Physical), 3)
vocal_vs_physical = round(cor(idol_school$Vocal, idol_school$Physical), 3)

# Use the cat function to output multiple lines at a time
cat(paste0("Correlation of vocal and dance:    ", vocal_vs_dance, "\n",
          "Correlation of dance and physical: ", dance_vs_physical, "\n",
          "Correlation of vocal and physical: ", vocal_vs_physical))
```

```
## Correlation of vocal and dance:    0.645
## Correlation of dance and physical: 0.509
## Correlation of vocal and physical: 0.664
```

Alternatively, we can also compute the correlation matrix for the three score categories (variables). The diagonal elements are always exactly 1 – because they represent the correlation of a variable with itself, which is a perfect positive correlation. The off-diagonal elements indicate the correlation coefficient between different categories.

```
round(cor(idol_school[,c("Vocal", "Dance", "Physical")]), 3)
```

```
##           Vocal Dance Physical
## Vocal    1.000 0.645    0.664
## Dance    0.645 1.000    0.509
## Physical 0.664 0.509    1.000
```

Create the scatterplots and/or correlation plots!

Use `ggplot` or not ?!

Need to explain the correlation coefficients and the K-pop context.

The training at a K-pop entertainment company in Korea usually includes vocal and dance lessons (Padget, 2017), so it is reasonable to see a high correlation between vocal and dance scores. Theoretically dance and physical should be highly correlated (Ngo et al., 2024), but in the Idol School dataset, we observed a slightly lower correlation in dance vs physical than in dance vs vocal. Physical strength is essential to dancing, but dance also includes other critical elements such as technique and aesthetic expression (Geukes et al., 2023).

Do more analysis to the Idol School data!

---

<sup>38</sup><https://bit.ly/431HHRS>



## 2.3 Idol School: Additional Resources

Students who were eliminated from the show:

[https://www.ptt.cc/bbs/fromis\\_9/M.1555819461.A.C73.html](https://www.ptt.cc/bbs/fromis_9/M.1555819461.A.C73.html)

In 2017, a Taiwanese blogger used random forests to predict the *Idol School* final ranking:

<https://shavid.pixnet.net/blog/post/331691281>

But this task turned out to be meaningless because the participant rankings were manipulated by the organizers.

*Idol School*: Vote Manipulation Investigation (2019)

<https://www.ptt.cc/bbs/KoreaStar/M.1624467107.A.D7F.html>

## 3 Read in the *Produce 48* Dataset

*Produce 48* dataset (2018)

Wikipedia data: [https://en.wikipedia.org/wiki/Produce\\_48](https://en.wikipedia.org/wiki/Produce_48)

Need to write the data description

*Produce 48* featured 96 contestants primarily from South Korea and Japan. The show was a collaboration between the Mnet's *Produce 101* series<sup>39</sup> and the Japanese *AKB48* idol group.<sup>40</sup>

Some former contestants in *Idol School* tried again in the *Produce 48* reality show in 2018.

A total of 12 contestants were eventually selected from *Produce 48* to create the time-limited girl group *IZ\*ONE*,<sup>41</sup> which was active during 2018-2021 in both Korea and Japan.

```
library(readxl)
produce_48_data = read_excel("UNFINISHED_Idol_School_Dataset.xlsx",
                           sheet="Produce_48_Dataset")

# Date of birth (DOB) should be date only, not a full timestamp.
produce_48_data$DOB = as.Date(produce_48_data$DOB)

columns_to_show = c("Name_Chn", "Name_Eng", "DOB",
                    "First_Eval", "Second_Eval", "Final_Rank")

produce_48_data[1:20, columns_to_show]
```

```
## # A tibble: 20 x 6
##   Name_Chn Name_Eng      DOB      First_Eval Second_Eval Final_Rank
##   <chr>    <chr>      <date>    <chr>        <chr>         <dbl>
## 1 張員瑛    Jang Won Young 2004-08-31 B           B             1
## 2 宮脇咲良  Miyawaki Sakura 1998-03-19 A           A             2
## 3 曹柔理    Jo Yuri        2001-10-22 A           F             3
## 4 崔叡娜    Choi Ye Na     1999-09-29 A           B             4
## 5 安俞真    An Yu Jin      2003-09-01 B           A             5
## 6 矢吹奈子  Yabuki Nako    2001-06-18 F           A             6
## 7 權恩妃    Kwon Eun Bi    1995-09-27 A           C             7
## 8 姜惠元    Kang Hye Won   1999-07-05 F           F             8
```

<sup>39</sup>[https://en.wikipedia.org/wiki/Produce\\_101](https://en.wikipedia.org/wiki/Produce_101)

<sup>40</sup><https://en.wikipedia.org/wiki/AKB48>

<sup>41</sup>[https://en.wikipedia.org/wiki/Iz\\*One](https://en.wikipedia.org/wiki/Iz*One)

##	9	本田仁美	Honda Hitomi	2001-10-06	C	A	9
##	10	金采源	Kim Chae Won	2000-08-01	B	B	10
##	11	金玟周	Kim Min Ju	2001-02-05	D	C	11
##	12	李彩演	Lee Chae Yeon	2000-01-11	A	A	12
##	13	韓霄瑗	Han Cho Won	2002-09-16	D	B	13
##	14	李佳恩	Lee Ka Eun	1994-08-20	A	A	14
##	15	宮崎美穂	Miyazaki Miho	1993-07-30	D	D	15
##	16	高橋朱里	Takahashi Juri	1997-10-03	B	A	16
##	17	竹内美宥	Takeuchi Miyu	1996-01-12	A	B	17
##	18	下尾美羽	Shitao Miu	2001-04-03	D	D	18
##	19	朴海允	Park Hae Yoon	1996-01-10	A	D	19
##	20	白間美瑠	Shiroma Miru	1997-10-14	B	D	20

Data entry complete for all contestants in *Produce 48*, including those who left in the middle of the show.

Create a matrix for the two sets of ratings.

For each rating, also check how many contestants are from Korea and how many are from Japan.

```
# UNFINISHED HERE
```

```
produce_48_data[81:96, columns_to_show]
```

```
## # A tibble: 16 x 6
##   Name_Chn   Name_Eng      DOB      First_Eval Second_Eval Final_Rank
##   <chr>      <chr>      <date>    <chr>      <chr>      <dbl>
## 1 克利絲汀   Alex Christine 1996-12-09 B          C          82
## 2 栗原紗英   Kurihara Sae   1996-06-20 F          D          83
## 3 趙英燕     Cho Yeong In   2001-10-31 B          C          84
## 4 淺井裕華   Asai Yuuka     2003-11-10 F          D          85
## 5 安藝媛     Ahn Ye Won     2001-02-10 F          F          86
## 6 內木志     Naiki Kokoro   1997-04-06 D          C          87
## 7 金有彬     Kim Yu Bin     2003-02-27 B          D          88
## 8 趙思朗     Cho Sa Rang    2003-09-05 B          F          89
## 9 崔韶恩     Choi So Eun    2001-09-19 B          C          90
## 10 篠崎彩奈   Shinozaki Ayana 1996-01-08 F          F          91
## 11 元書妍     Won Seo Yeon   2000-05-23 C          F          92
## 12 月足天音   Tsukiashi Amane 1999-10-26 F          F          100
## 13 田中美久   Tanaka Miku    2001-09-12 F          C          100
## 14 梅山戀和   Umeyama Kokona 2003-08-07 F          X          100
## 15 植村梓     Uemura Azusa   1999-02-04 F          X          100
## 16 松井珠理奈 Matsui Jurina   1997-03-08 B          B          100
```

Let's look at the nationality breakdown of *Produce 48* contestants. Although *Produce 48* advertised a collaboration between Korean and Japanese entertainment groups, the Korea-Japan split is not 1-1 among participants. The majority (56%) of contestants are domestic within South Korea, while a lower but remarkable (40%) proportion is from Japan. Also, two contestants are from China and another one is from the United States.

```
table(produce_48_data$Country)
```

```
##
## China Japan Korea  USA
##    2    39    54    1
```

First Evaluation

```
# UNFINISHED HERE
table(produce_48_data$First_Eval, dnn="First_Eval")
```

```
## First_Eval
## A B C D F
## 15 25 22 15 19
```

Second Evaluation

```
# UNFINISHED HERE
table(produce_48_data$Second_Eval, dnn="Second_Eval")
```

```
## Second_Eval
## A B C D F X
## 14 20 22 16 22 2
```

Cross-table: **First\_Eval** as row, and **Country** as column

Observation: The majority of contestants from Japan were placed in D or F (lowest grades), while very few contestants from Korea did.

Idol training differs in Korea and Japan (citation needed)

Korea - primary focus on vocal and dance techniques

Japan - artistic interpretation ?!

```
table(produce_48_data$First_Eval, produce_48_data$Country,
      dnn=c("First_Eval", "Country"))
```

```
##           Country
## First_Eval China Japan Korea USA
##           A      0      2     13   0
##           B      2      4     18   1
##           C      0      5     17   0
##           D      0     11      4   0
##           F      0     17      2   0
```

Cross-table: **Second\_Eval** as row, and **Country** as column

Observation: Contestants from Japan faced a little better in the second evaluation – approximately half of them received a satisfactory grade (A, B, or C).

The scoring of domestic contestants in Korean became harsher this time – one third of them were rated unsatisfactory (D or F).

Need to specify the (inferred) reason

```
table(produce_48_data$Second_Eval, produce_48_data$Country,
      dnn=c("Second_Eval", "Country"))
```

```
##           Country
## Second_Eval China Japan Korea USA
```

```
##           A      0      4     10     0
##           B      0      6     14     0
##           C      1      8     12     1
##           D      1      9      6     0
##           F      0     10     12     0
##           X      0      2      0     0
```

Cross-table: **First\_Eval** as row, and **Second\_Eval** as column

List the names of the six contestants who got A  $\rightarrow$  A.

Jo Yuri (曹柔理): A  $\rightarrow$  F

Yabuki Nako (矢吹奈子): F  $\rightarrow$  A

What about other participants?

```
table(produce_48_data$First_Eval, produce_48_data$Second_Eval,
      dnn=c("First_Eval", "Second_Eval"))
```

```
##           Second_Eval
## First_Eval A B C D F X
##           A 6 3 4 1 1 0
##           B 4 8 5 5 3 0
##           C 3 6 4 3 6 0
##           D 0 3 5 3 4 0
##           F 1 0 4 4 8 2
```

Next step: Breakdown the 1st-cross-2nd table by country (Korea and Japan)

1st-cross-2nd table: Korea

```
produce_48_korea = produce_48_data[which(produce_48_data$Country=="Korea"),]
table(produce_48_korea$First_Eval, produce_48_korea$Second_Eval,
      dnn=c("First_Eval", "Second_Eval"))
```

```
##           Second_Eval
## First_Eval A B C D F
##           A 5 2 4 1 1
##           B 3 7 3 2 3
##           C 2 4 4 3 4
##           D 0 1 1 0 2
##           F 0 0 0 0 2
```

1st-cross-2nd table: Japan

```
produce_48_japan = produce_48_data[which(produce_48_data$Country=="Japan"),]
table(produce_48_japan$First_Eval, produce_48_japan$Second_Eval,
      dnn=c("First_Eval", "Second_Eval"))
```

```
##           Second_Eval
## First_Eval A B C D F X
##           A 1 1 0 0 0 0
##           B 1 1 0 2 0 0
##           C 1 2 0 0 2 0
##           D 0 2 4 3 2 0
##           F 1 0 4 4 6 2
```

## 4 Tentative Placeholders

Write something here

### 4.1 Test for Non-English Characters

CJK = Chinese, Japanese, Korean

Chinese example

RStudio 有辦法打中文嗎？

```
print(" 大家好，很高興能認識你們！")
```

```
## [1] "大家好，很高興能認識你們！"
```

Japanese example

思い出にするにはまだ早すぎる

```
print(" みやわき さくら")
```

```
## [1] "みやわき さくら"
```

```
print(" 宮脇 咲良")
```

```
## [1] "宮脇 咲良"
```

This template does not support Korean characters yet.

### 4.2 R Markdown Narrative

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed      dist
##  Min.   : 4.0    Min.   :  2.00
##  1st Qu.:12.0    1st Qu.: 26.00
##  Median :15.0    Median : 36.00
##  Mean   :15.4    Mean   : 42.98
##  3rd Qu.:19.0    3rd Qu.: 56.00
##  Max.   :25.0    Max.   :120.00
```

### 4.3 Including Plots

You can also embed plots, for example in Figure 1:

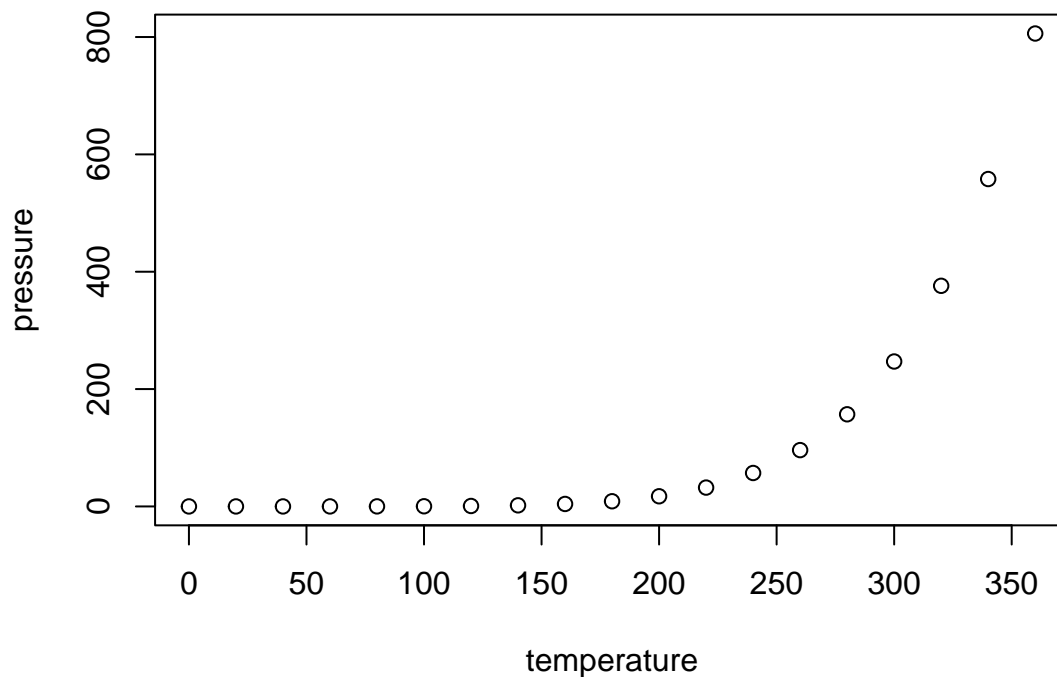


Figure 1: Test Plot

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

## Acknowledgments

The author is immensely grateful to her significant other, Hugh Hendrickson, for providing his support in the author's professional development.

Technical discussions: Cheng-Shun Liu (劉承順) and Chih-Kuang Lee (李治廣, Kevin).

## References

- Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2024). *rmarkdown: Dynamic Documents for R*. R package version 2.29. <https://github.com/rstudio/rmarkdown>.
- Chai, C. P. (2024). Statistical analysis of high school and college entrance exam scores in Taiwan with online data. *Preprint on ResearchGate*. <http://dx.doi.org/10.13140/RG.2.2.29468.91520/1>.
- Chen, I. (2023). Expansion of K-pop in the global market. *American Journal of Student Research*, 1(1):1–6. <https://doi.org/10.70251/hyjr2348.1115>.
- Cho, J., Bian, Y., and Lee, J. (2023). Leading digital business model transformation in the K-pop industry: The case of SM Entertainment. *Asia Pacific Business Review*, 29(5):1394–1424. <https://doi.org/10.1080/13602381.2023.2229761>.

- Geukes, K., Hecht, V., Utesch, T., Bläsing, B., and Back, M. (2023). Mirror, mirror on the wall, who is the fairest dancer of them all? A naturalistic lens model study on the judgment of dance performance. *Psychology of Sport and Exercise*, 67:102436. <https://doi.org/10.1016/j.psychsport.2023.102436>.
- Han, C. and Pothong, A. (2021). K-pop's ingredients of success. *Journal of Student Research*, 10(2):1–11. <https://doi.org/10.47611/jsrhs.v10i2.1431>.
- Kang, J. M. (2017). Rediscovering the idols: K-pop idols behind the mask. *Celebrity Studies*, 8(1):136–141. <https://doi.org/10.1080/19392397.2016.1272859>.
- Khium, L. K. (2013). K-Pop dance trackers and cover dancers: Global cosmopolitanization and local spatialization. In *The Korean Wave*, pages 165–181. Routledge, London, United Kingdom. <https://doi.org/10.4324/9781315859064>.
- Kim, J. and Kwon, S.-H. (2022). K-pop's global success and its innovative production system. *Sustainability*, 14(17):11101. <https://doi.org/10.3390/su141711101>.
- Kim, J.-h., Jung, S.-h., Roh, J.-s., and Choi, H.-j. (2021). Success factors and sustainability of the K-pop industry: A structural equation model and fuzzy set analysis. *Sustainability*, 13(11):5927. <https://doi.org/10.3390/su13115927>.
- Kim, J.-M. (2020). The linguistics of name translation: Preferred personal and business names in English, Korean, and Chinese. *Names: A Journal of Onomastics*, 68(2):104–124. <https://doi.org/10.1080/00277738.2020.1731242>.
- Lee, G. T. (2024). The establishment of K-Pop: K-Pop's main characteristics. In *The Palgrave Handbook of Critical Music Industry Studies*, pages 293–313. Springer, Cham, Switzerland. [https://doi.org/10.1007/978-3-031-64013-1\\_18](https://doi.org/10.1007/978-3-031-64013-1_18).
- Lee, H. J. and Jin, K. Y. Y. (2019). *K-Pop Idols: Popular Culture and the Emergence of the Korean Music Industry*. Lexington Books, Lanham MD, United States. <https://rowman.com/isbn/9781498588263>.
- Lee, H.-K. and Zhang, X. (2021). The Korean wave as a source of implicit cultural policy: Making of a neoliberal subjectivity in a Korean style. *International Journal of Cultural Studies*, 24(3):521–537. <https://doi.org/10.1177/1367877920961108>.
- Liu, J. (2025). K-pop's overcrowded market: Analyzing the effects of excessive debuts on industry revenue and growth. *Law and Economy*, 4(2):32–43. <https://www.paradigmexpress.org/le/article/view/1544>.
- Lockwood, D. (2021). *Fooled by the winners: How survivor bias deceives us*. Greenleaf Book Group, Austin TX, United States. <https://greenleafbookgroup.com/titles/fooled-by-the-winners>.
- Malik, T. H. (2023). K-pop music diffusion in Korea and East Asia: The convergence of visual technology and concrete narratives. *Asia Pacific Business Review*, 29(5):1251–1274. <https://doi.org/10.1080/13602381.2023.2237908>.
- Min, B.-S. (2024). The K-pop industry: Competitiveness and sustainability. *International Journal of Cultural Policy*, pages 1–17. <https://doi.org/10.1080/10286632.2024.2366979>.
- Miroudot, S. (2024). What's behind the 'K'? Common audio features of Korean popular music before and after the rise of K-pop. *Popular Music*, 0(0):1–22. <https://doi.org/10.1017/S0261143024000187>.
- Ngo, J. K., Lu, J., Cloak, R., Wong, D. P., Devonport, T., and Wyon, M. A. (2024). Strength and conditioning in dance: A systematic review and meta-analysis. *European Journal of Sport Science*, 24(6):637–652. <https://doi.org/10.1002/ejsc.12111>.
- Padget, F. (2017). What are the difficulties of being a Korean pop idol and to what extent do they outweigh the benefits? *San Francisco, CA: Academia. edu-Share Research*.

- Sun, Y. (2022). Identifying the factors leading to the globalization of K-Pop. In *2022 International Conference on Science Education and Art Appreciation (SEAA 2022)*, pages 769–776, Amsterdam, Netherlands. Atlantis Press (part of Springer Nature). [https://doi.org/10.2991/978-2-494069-05-3\\_94](https://doi.org/10.2991/978-2-494069-05-3_94).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York NY, United States. <https://ggplot2.tidyverse.org>.
- Wickham, H. and Bryan, J. (2023). *readxl: Read Excel Files*. R package version 1.4.3. <https://CRAN.R-project.org/package=readxl>.
- Yoshimitsu, M. (2020). Affective economics in the East Asian media and entertainment industry: Comparative case studies of music competition television series. *Review of East Asian Affairs*, 3(12). Institute of East Asian Studies, University of Nagasaki. [http://54.64.218.135/dspace/bitstream/10561/1602/1/v12p83\\_yoshimitsu.pdf](http://54.64.218.135/dspace/bitstream/10561/1602/1/v12p83_yoshimitsu.pdf).