

# K-Pop Data Analysis

Christine P. Chai  
cpchai21@gmail.com

January 17, 2025

Starting in 2024.

Test citation (Chai, 2024)

## Executive Summary

Write something here

## Disclaimer

This manuscript is written solely by the author using R Markdown (Allaire et al., 2024).<sup>1</sup> It is not created 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

The author became interested in K-Pop music (Korean popular music) from the debut of Tzuyu (Chou Tzu-Yu, 周子瑜).<sup>2</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>3</sup> and eventually got added to the newly-formed girl group *TWICE*.<sup>4</sup>

Later that year, ...

Describe the flag controversy incident.<sup>5</sup>

(a lot more content here)

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

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

Started with the *Produce X 101* (2019)

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

---

<sup>1</sup><https://rmarkdown.rstudio.com/>

<sup>2</sup><https://en.wikipedia.org/wiki/Tzuyu>

<sup>3</sup>[https://en.wikipedia.org/wiki/Sixteen\\_\(TV\\_program\)](https://en.wikipedia.org/wiki/Sixteen_(TV_program))

<sup>4</sup><https://en.wikipedia.org/wiki/Twice>

<sup>5</sup><https://bit.ly/3DOcNIP>

The mysterious 29978 number in *Produce X 101*:

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

Mnet admitted to manipulating the votes in the *Produce 101* series and the subsequent reality shows, including *Idol School*.

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

*Idol School*: Vote Manipulation Investigation (2019)

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

## 1.1 Read in the *Idol School* Dataset

*Idol School* (偶像學校) (2017)

Motivation: One of the contestants, Snowbaby (蔡瑞雪),<sup>6</sup> is also from Taiwan. In fact, Snowbaby<sup>7</sup> graduated from Taipei First Girls' High School,<sup>8</sup> the same high school as the author did.

In the live reality show *Idol School*, nine winners were selected to form the girl group *fromis\_9*.<sup>9</sup> This girl group debuted in 2018 and remained active until the contract ended in 2024.

### 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)

We manually copy-pasted the contestant data from Wikipedia into an Excel workbook (.xlsx), and used the R package `readxl` (Wickham and Bryan, 2023) to load the dataset.

### Why didn't we store the dataset in .csv format?

For Chinese characters and having multiple data sheets in the same Excel file for consolidation.

Since the English 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.

```
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[1:20, columns_to_show]
```

```
## # A tibble: 20 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
```

<sup>6</sup>Snowbaby's YouTube channel: <https://www.youtube.com/@snowbaby>

<sup>7</sup><https://bit.ly/424u3gv>

<sup>8</sup><https://www.fg.tp.edu.tw/>

<sup>9</sup>[https://en.wikipedia.org/wiki/Fromis\\_9](https://en.wikipedia.org/wiki/Fromis_9)

##	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
##	11	盧知宣	Roh Ji Sun	1998-11-23	6.5	7	6.5	6.67
##	12	裴恩英	Bae Eun Yeong	1997-05-23	7	9.3	3.5	6.6
##	13	朴池原	Park Ji Won	1998-03-20	7.9	5	6.2	6.37
##	14	曹侑彬	Cho Yu Bin	1999-10-09	5.9	9	4	6.3
##	15	李賽綸	Lee Sae Rom	1997-01-07	5	5.1	8.7	6.27
##	16	秋元喜	Chu Won Hui	1999-04-14	5.7	7.4	5	6.03
##	17	李多熙	Lee Da Hee	1996-04-25	6.4	4.9	4.9	5.4
##	18	賓荷娜	Sky / Bin Ha Neul	1999-12-14	4	5.4	6.1	5.17
##	19	李瑞淵	Lee Seo Yeon	2000-01-22	6.1	6.3	2	4.8
##	20	楊璉智	Yang Yeon Ji	1996-01-03	4.9	7.5	1.6	4.67

## 1.2 *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 each category.

**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 by the physical scores, we found two 3.5's and two 1.2's. 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.

**Physical:** We found two 3.5's and two 1.2's after sorting the scores.

In the video clip, Park Ji Won (朴池原) and her partner were the first runner-up in the group physical exercise.<sup>10</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>11</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>12</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.

**Physical:** We found additional two 1.2's after sorting the scores.

The two 1.2 scores are more difficult to check for the underlying values, probably because they occurred in two contestants of lower ranking.<sup>13</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>14</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.

<sup>10</sup>Screenshot of the group physical exercise: <https://bit.ly/4a7QT9m>

<sup>11</sup><https://bit.ly/400KUHH>

<sup>12</sup>Park Ji Won was the main vocalist in *fromis\_9*. <https://bit.ly/402yCFI>

<sup>13</sup>Physical scores of all contestants in *Idol School*: <https://bit.ly/3DRNK0Z>

<sup>14</sup><https://translate.google.com/>

*Idol School* (2017): Videos with subtitles in Simplified Chinese are available on the Bilibili platform.<sup>15</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

```
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)

sorted_scores_df[1:10,]
```

##	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

Check for the mean and median of each category score

```
# UNFINISHED HERE

# Output a table for the mean and median for (vocal, dance, physical)

# Columns: Vocal, Dance, Physical
# Rows: Mean, Median

# Examples:
# mean(idol_school$Dance) # 5.35122
# median(idol_school$Dance) # 5.5

# Rounding to two decimal places?!
```

Correlation matrix

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

Diagonal elements are always exactly 1.

Create the scatterplots and/or correlation plots!

Use **ggplot** or not ?!

Why is the correlation so low between dance and physical scores?

---

<sup>15</sup><https://www.bilibili.com/video/BV1554y1C7wj/>

Dance is mainly about technique, not always about the person's physical ability. (citation needed)

Contestants with a remarkably high score in dance but a low score in physical:

e.g. Bae Eun Yeong (裴恩英)

e.g. Lee Hae In (李海印)

```
# UNFINISHED HERE
cor(idol_school[,c("Vocal", "Dance", "Physical"])]
```

```
##           Vocal      Dance  Physical
## Vocal      1.0000000 0.6821046 0.6834680
## Dance      0.6821046 1.0000000 0.5426207
## Physical   0.6834680 0.5426207 1.0000000
```

Alternatively, we can also obtain the pairwise correlation of each category.

```
# UNFINISHED HERE
cor(idol_school$Vocal, idol_school$Dance)
```

```
## [1] 0.6821046
```

```
# UNFINISHED HERE
# Need to print all three pairs.
# cor(idol_school$Dance, idol_school$Physical)
# cor(idol_school$Vocal, idol_school$Physical)
```

### 1.3 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)

Someone else used random forests to predict the final ranking:

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

### 1.4 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)

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>16</sup> which was active during 2018-2021 in both Korea and Japan.

```
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)
```

---

<sup>16</sup>[https://en.wikipedia.org/wiki/Iz\\*One](https://en.wikipedia.org/wiki/Iz*One)

```
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
## 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 宮崎美穗 Miyazaki Miho  1993-07-30 D          D          15
## 16 高橋朱里 Takahashi Juri  1997-10-03 B          A          16
## 17 竹內美宥 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
```

Still working on the data entry.

```
# UNFINISHED HERE
produce_48_data[51:60, columns_to_show]
```

```
## # A tibble: 10 x 6
##   Name_Chn Name_Eng      DOB      First_Eval Second_Eval Final_Rank
##   <chr>    <chr>      <date>      <chr>      <chr>      <dbl>
## 1 李宥姪    Lee Yu Jeong 2004-06-14 C          C          51
## 2 本村碧唯 Motomura Aoi  1997-05-31 D          B          52
## 3 朴珉志    <NA>        NA          <NA>      <NA>      53
## 4 柳旻瑛    <NA>        NA          <NA>      <NA>      54
## 5 朴書英    <NA>        NA          <NA>      <NA>      55
## 6 王珂      Wang Ke      2000-11-05 B          D          56
## 7 曹佳賢    <NA>        NA          <NA>      <NA>      57
## 8 中野郁海 Nakano Ikumi  2000-08-20 D          B          59
## 9 黃召硯    <NA>        NA          <NA>      <NA>      60
## 10 申帥賢    <NA>        NA          <NA>      <NA>      61
```

## 2 Tentative Placeholders

Write something here

## 2.1 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.

## 2.2 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
```

## 2.3 Including Plots

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

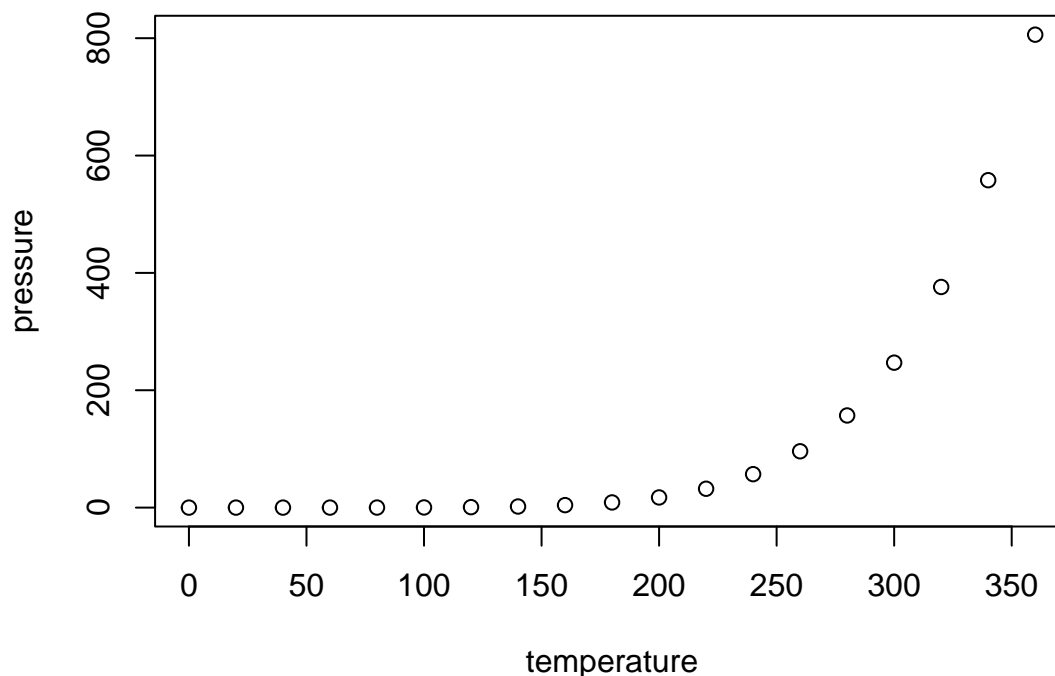


Figure 1: Test Plot

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 (Adson) Liu and Chih-Kuang (Kevin) Lee.

## 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>.
- 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>.
- Wickham, H. and Bryan, J. (2023). *readxl: Read Excel Files*. R package version 1.4.3. <https://CRAN.R-project.org/package=readxl>.