K-Pop Data Analysis

Christine P. Chai
cpchai21@gmail.com

January 7, 2025

Starting in 2024.

Test citation (Chai, 2024)

Executive Summary

Write something here

Disclaimer

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

How the author got interested in K-Pop music (Korean popular music):

Tzuyu (Chou Tzu-Yu, 周子瑜)[1]

(a lot more content here)

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

1.1 Read in the Idol School Dataset

Idol School (偶像學校) (2017)

Motivation: One of the contestants, Snowbaby (蔡瑞雪),[2] is originally from Taiwan. In fact, Snowbaby[3] graduated from Taipei First Girls' High School,[4] the same high school as the author did.

Need to write the data description

Wikipedia data: https://en.wikipedia.org/wiki/List_of_Idol_School_contestants

Why did we include the date of birth (DOB) of each contestant?

[1] https://en.wikipedia.org/wiki/Tzuyu
[2] Snowbaby's YouTube channel: https://www.youtube.com/@snowbaby
[3] https://bit.ly/424u3gv
[4] https://www.fg.tp.edu.tw/

```
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", "Ability_Rank")

idol_school[1:10, columns_to_show]
```

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

1.2 Idol School: Exploratory Data Analysis

What changes did we make from the Wikipedia data?

Our presumption: In each category, no two contestants should have the same score.

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

The two 3.5 scores belong to adjacent cells in the Wikipedia data.

Physical testing contains a group exercise and an individual exercise.

In the video clip, Park Ji Won (朴池原) and her partner were the first runner-up in the group exercise.[5] 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,[6] Ji Won's physical score should be 6.2.

The Wikipedia table shows an inconsistency in the overall score, i.e., the average across the three categories.

Ji Won's vocal score was 7.9, and her dance was 5. These scores seem to be reasonable for Ji Won, because she is known for excellent singing and decent dancing as a performer.[7] 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.

[5] Screenshot of the group physical exercise: https://bit.ly/4a7QT9m
[6] https://bit.ly/400KUhH
[7] Park Ji Won is the main vocalist in Fromis 9. https://bit.ly/402yCFI

The second scenario is more likely.
Evidence we found in the video clip.

The two 1.2 scores are more difficult to check for the underlying values.

Especially that they occurred in two contestants with lower ranking.[8]

With the help of Google Translate:[9]
Can translate Korean text in an image back to English text.
Finally, we discovered that Michelle White (懷特·米雪兒)'s physical score should be 1.3, not 1.2.

Idol School (2017): Videos with subtitles in Simplified Chinese
https://www.bilibili.com/video/BV1554y1C7wj/

Screenshots saved:
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?!
```

[8] Physical scores of all contestants in Idol School: https://bit.ly/3DRNK0Z
[9] https://translate.google.com/

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

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)

```
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)

# UNFINISHED HERE:
# Decide on which columns and rows to show here.

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 宮崎美穂 Miyazaki Miho   1993-07-30 D          D                   15
## 16 <NA>     <NA>            NA         <NA>       <NA>                16
## 17 <NA>     <NA>            NA         <NA>       <NA>                17
## 18 <NA>     <NA>            NA         <NA>       <NA>                18
## 19 朴海允   Park Hae Yoon   1996-01-10 A          D                   19
## 20 <NA>     <NA>            NA         <NA>       <NA>                20
```

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:

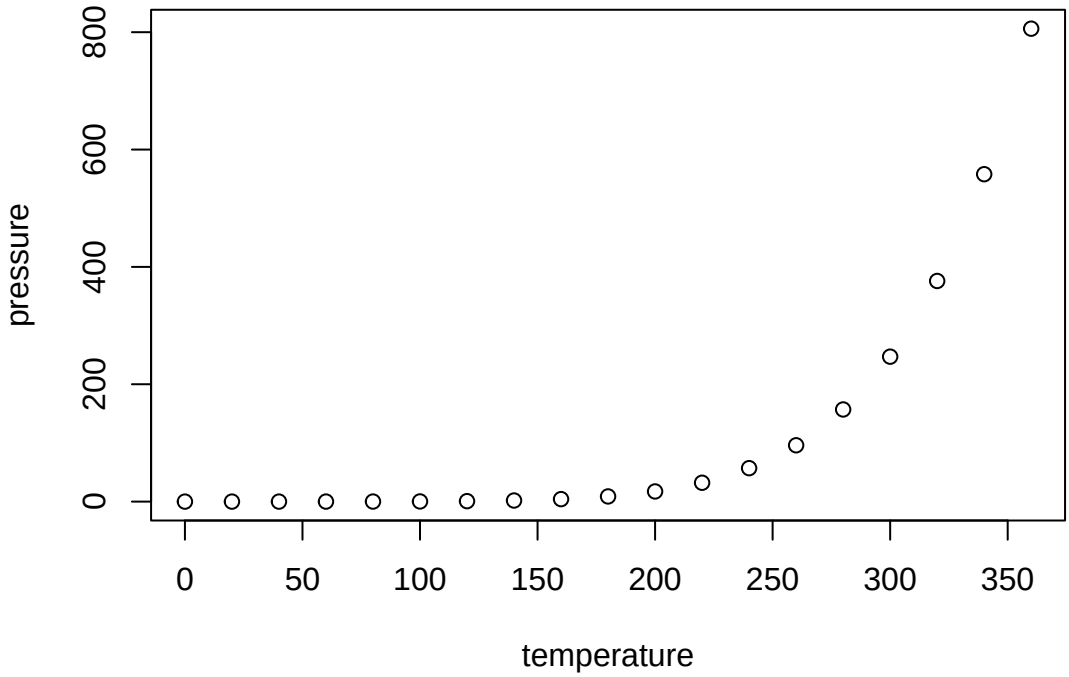

Figure 1: Test Plot

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

Acknowledgments

Write something here

References

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.