K-Pop Data Analysis

Christine P. Chai
cpchai21@gmail.com

January 3, 2025

Starting in 2024.

Test citation (Chai, 2024)

1 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.

2 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.

2.1 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.

Need to write the data description

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

[1] https://en.wikipedia.org/wiki/Tzuyu
[2] Snowbaby's YouTube channel: https://www.youtube.com/@snowbaby

```r
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
```

2.2 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. We were surprised that Park Ji Won (朴池原)'s physical score was only 3.5.

According to the physical score table in the video screenshots, Park Ji Won (朴池原)'s physical score should be 6.2.
The overall score, i.e., the average across the three categories => inconsistent.

The two 1.2 scores are more difficult to check for the underlying values.
With the help of Google Translate:[3]
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

[3] https://translate.google.com/

```r
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
# Make the cbind object a data.frame!
# cbind(vocal_sorted, dance_sorted, physical_sorted)
```

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

Someone else used random forests to predict the final ranking:
https://shavid.pixnet.net/blog/post/331691281

2.4 Read in the Produce 48 Dataset

Produce 48 dataset (2018)

```r
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:
# Decide on which columns and rows to show here.

produce_48_data
```

```
## # A tibble: 20 x 9
##    Name_Chn Name_Eng        DOB        First_Eval Second_Eval Country Final_Rank
##    <chr>    <chr>           <date>     <chr>      <chr>       <chr>         <dbl>
##  1 張員瑛   Jang Won Young  2004-08-31 B          B           Korea             1
##  2 宮脇咲良 Miyawaki Sakura 1998-03-19 A          A           Japan             2
##  3 曹柔理   Jo Yuri         2001-10-22 A          F           Korea             3
##  4 <NA>     <NA>            NA         <NA>       <NA>        Korea             4
##  5 <NA>     <NA>            NA         <NA>       <NA>        Korea             5
##  6 矢吹奈子 Yabuki Nako     2001-06-18 F          A           Japan             6
##  7 <NA>     <NA>            NA         <NA>       <NA>        Korea             7
##  8 <NA>     <NA>            NA         <NA>       <NA>        Korea             8
##  9 <NA>     <NA>            NA         <NA>       <NA>        Japan             9
## 10 <NA>     <NA>            NA         <NA>       <NA>        Korea            10
## 11 <NA>     <NA>            NA         <NA>       <NA>        Korea            11
## 12 <NA>     <NA>            NA         <NA>       <NA>        Korea            12
## 13 <NA>     <NA>            NA         <NA>       <NA>        Korea            13
## 14 <NA>     <NA>            NA         <NA>       <NA>        Korea            14
## 15 <NA>     <NA>            NA         <NA>       <NA>        <NA>             15
## 16 <NA>     <NA>            NA         <NA>       <NA>        <NA>             16
## 17 <NA>     <NA>            NA         <NA>       <NA>        <NA>             17
## 18 <NA>     <NA>            NA         <NA>       <NA>        <NA>             18
```

```
## 19 <NA>     <NA>            NA         <NA>       <NA>       <NA>           19
## 20 <NA>     <NA>            NA         <NA>       <NA>       <NA>           20
##    Round_Eliminated Special_Notes
##    <chr>            <lgl>
##  1 Survived         NA
##  2 Survived         NA
##  3 Survived         NA
##  4 Survived         NA
##  5 Survived         NA
##  6 Survived         NA
##  7 Survived         NA
##  8 Survived         NA
##  9 Survived         NA
## 10 Survived         NA
## 11 Survived         NA
## 12 Survived         NA
## 13 <NA>             NA
## 14 <NA>             NA
## 15 <NA>             NA
## 16 <NA>             NA
## 17 <NA>             NA
## 18 <NA>             NA
## 19 <NA>             NA
## 20 <NA>             NA
```

3 Tentative Placeholders

Write something here

3.1 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.

3.2 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.

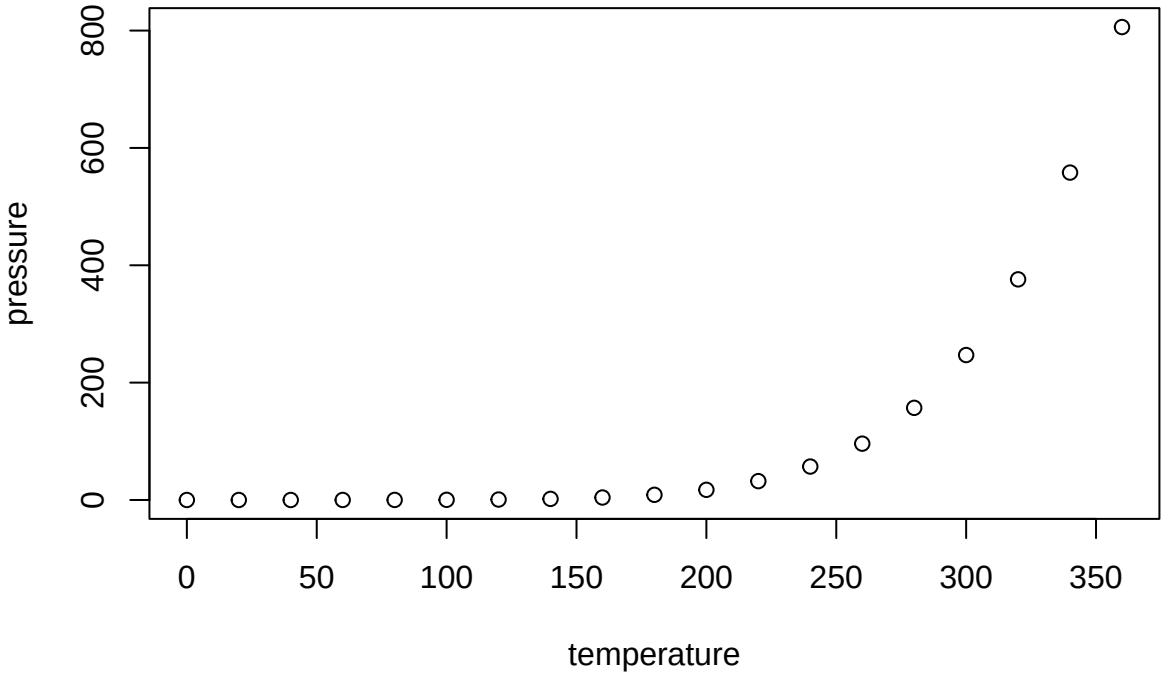
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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
```

3.3 Including Plots

You can also embed plots, for example:



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.