## I. Introduction and Research Interests

Although most of the people nowadays, especially the younger generations, no longer use CD players when listening to music, physical album sales is still a big portion that help made up the revenue for K-pop groups and their agencies. This is to say, purchasing physical albums is less than listening to the CDs but more of showing a fan's support to the artist beyond streaming that hardly cost any money except for the subscription to the platform. This research was done by Feier Su, which aimed to identify which driving factors correlated most strongly with high physical album sales in the context of fourth generation K-pop groups up to the end of November, 2025.

## II. Data Collection

Since this research is about examining the strength of correlation between predicting factors and the outcome, it is necessary to first set up a dependent variable and a series of independent variables based on hypotheses and research interest.

The dependent variable was the *physical album sales* and the data was from Soridata where it contains all K-pop groups' statistics and its main source is Circle Chart, a Korean ranking system similar to Billboard. This link gathered sales of K-pop groups of all generations. For this study, only top 20 K-pop groups from the fourth generation are the target groups due to the ease of controlling variables that they were debuted around the same time during COVID-19 and the deadline ensured that all groups had a comparable amount of time to release music, build fandoms, and accumulate sales.

On the other hand, 8 independent variables and 160 data points in total were collected, including 2 dummy variables and 6 numerical variables. Similar to the approach done for the dependent variable, the script needed to filter out all groups that are not the target groups.

1. *Group type*, or the gender of the group (boy group or girl group)

   *H1: boy groups are more likely to generate higher sales*

2. *Company label* (SM, JYP, YG, HYBE, etc.)

   *H2: big and leading companies such as SM and Hybe tend to generate higher sales*

These two dummy variables were simply scraped from Wikipedia files of each group.

3. *Digital score* of each group via Soridata

   *H3: higher digital scores would result in higher sales*

   A weighted indicator provided by Circle Chart and sorted by Soridata that calculated a group's digital streaming and download performance through a covert formula.

4. *YouTube total views* of each group via kworb.net

   *H4: higher views would result in higher sales*

5. *Number of YouTube channel subscribers* via Kpopping

   *H5: more subscribers would result in higher sales*

6. *Number of annual awards* via Soridata

   *H6: more annual awards would result in higher sales*

   These are year-end major industry ceremonies that celebrate achievements and give awards to the most successful artists, songs, and albums in the Korean music industry based on sales, streaming, voting, and expert evaluations.

7. *Number of music show awards* via Kpopping

   *H7: more music show awards would result in higher sales*

These are weekly South Korean music shows such as Music Bank, Inkigayo, M Countdown, Show Champion, Music Core, and The Show feature idols performing their latest songs and competing for weekly awards determined by sales, streams, votes, and broadcast metrics. These programs act as essential promotional stages for newly released music.

8. *Days to first win* via Kpopping

*H8: less days to first win would result in higher sales*

This refers to the number of days it took for each K-pop group to earn their first win on a music program

Changes from the original proposal included narrowing down the sample size where the initial proposal aimed to gather top 100 K-pop groups from all generations; however, group debuted in different generations may have disparate marketing approaches and the different durations since they debuted makes it incomparable when calculating album sales. The second change was replacing the original method of scraping YouTube views, subscriber counts, annual awards, and music show wins from the initially used websites with data from alternative statistical sources due to failure to detect the table format on Soridata and that Soridata has an anti-scraping measure that prevents too much scraping attempts. The third change was eliminating one of the hypothesized predicting factors, number of concert tours, as it lacks availability of accurate data.

## III. Results and Discussion

The first test was a bar plot showing correlation significance by p-values and correlation coefficients of the numerical variables with physical album sales. The bars were also colored to

indicate that dark green meant the most strong and significant correlation with $p < .001$, the lightest green color meant moderately significant correlation with $p < .1$, and gray bars meant no significance. The plot suggested that among all 6 numerical variables, number of YouTube subscribers, YouTube total views, annual awards, and music show awards were significantly and positively correlated with physical sales, where YouTube statistics were the strongest predictors. In other words, digital presence and industry recognition play key roles in boosting physical sales, guiding marketing efforts to focus more on YouTube engagement and award participation for better commercial success.

The second visualization was a bar graph that ranked the total sales by companies, which it displayed the overall commercial success and market share of a company label although it may simply due to the fact that big companies with more groups tend to generate more sales. The third visualization was a similar bar graph that ranked the companies but replaced total sales by median sales, which also helped reveal the competitive dynamics between companies. To summarize the results, HYBE led the K-pop industry in both total physical album sales and median sales per group, showing strong overall market dominance and stable sales performance. JYP ranked second in total sales but fell to sixth in median sales, indicating reliance on a few phenomenal groups rather than consistent average performance across all groups. Cube and Starship show strong median sales, reflecting balanced and stable sales among their groups.

Last, an ANOVA with post-hoc test together with a generation of box plot was run to examine whether if company labels help create differences when making a success in physical album sales. Although the box plot demonstrated that there was no overlap between leading and lagging companies, the p-value ($p = .913$) of the ANOVA test as well as the post-hoc test results suggested that the differences were not significant; however, this can also be caused by the small

data size. Next, the same approach was done to test if there was a difference between group type (boy or girl) in generating sales. The results also showed that there was no significant difference between genders ($p$ = .154); however, the data distribution showed that boy groups generally earned higher revenues. In sum, these findings provided valuable insights for stakeholders interested in market positioning, artist development, and investment strategies within the K-pop industry.

## IV. Limitations and Future Work

For future research, it should consider expanding the sample size to include more groups across multiple generations with appropriate controls of multicollinearity to enhance the generalizability and statistical power of the analysis. It should also consider incorporating more diverse data variables and sources such as social media engagement beyond YouTube, international market reach, concert tours, and merchandising revenue, to capture a more thorough picture of commercial success. Furthermore, employing advanced modeling techniques such as multivariate regression or machine learning could also better isolate the effects of various factors and explore more complex relationships. Finally, in addition to quantitative findings, qualitative insights from fan surveys would also provide valuable context, though not being the focus of this study.