K-Pop Group Success Factors: A Data-Driven Look at Company Power and Generational Differences

Introduction

Since the first widely recognized K-pop idol group debuted in 1992, the industry has grown from a small collection of pioneering acts into a global cultural force. Throughout the first and second generations of K-pop, spanning roughly 1992 to 2010, the number of new groups debuting each year remained relatively low, rarely reaching double digits. Starting in the third generation, however, the pace of debuts increased sharply. Since around 2012 to now, it has become very common for 50 or more new groups to debut each year, reflecting the industry's immense growth. As a result of this growth, however, competition has become increasingly fierce, with recent years reporting 70 or more groups debuting yearly. Out of those groups, though, only a handful of five to seven of them reach the public eye, with an even smaller number of them achieving global fame.

This raises an important question: What is it that determines whether a newly debuted K-pop group will succeed or not?

With some K-pop idols training for as long as eight years and others debuting after only a few months, these artists invest their entire lives into an industry that offers almost no guarantee of success. This raises the question of why certain groups are discovered and embraced by the public right away, while others, often called "nugu groups," meaning groups that are basically unknown, can spend years releasing music before they finally get a hit and become recognized.

As both a data science student and a longtime K-pop listener, I wanted to explore whether success is primarily shaped by talent and music quality or whether it is more strongly influenced by structural factors such as company resources, debut timing, and group demographics, with a strong emphasis on company resources. Using a dataset that includes company information, debut dates, group type, digital views, album sales, and music show wins, this project investigates how predictable K-pop success actually is. I also compare objective measures of success with my own familiarity as a fan to see where personal perception aligns with or diverges from industry outcomes. With this project, I am ultimately attempting to reveal how uneven the landscape of K-pop can be, as long before a group even releases its first song, certain advantages or disadvantages may already shape its future trajectory.

Although I'm using this dataset to explore patterns in K-pop success, I also recognize that the data has a lot of limitations. It only includes a small sample of groups; many important variables are missing, and the measures of success available are far from perfect. My conclusions are based on what I can analyze with the data I have, not a complete picture of the industry. I go into these limitations more fully in the discussion section, but it's important to keep in mind from the start that this project is more exploratory than definitive.

Data & Methods

For this project, I used a publicly available K-pop debut dataset from Kaggle that includes information on around 120 different groups. Each row represents one group and contains details such as the company they debuted under, their debut date, their group type/gender, and several indicators of how well they performed over time, including physical album sales, YouTube view counts, and the number of music show awards they received. The original dataset also included astrology-based features, but I removed all of those. I wanted this project to focus on concrete, measurable factors rather than symbolic traits that do not fit the type of analysis I am trying to do.

After cleaning up the dataset, I added a personal "knowledge score" column that I planned to use later. This score ranged from zero to two and simply captured how well I personally knew each group. My goal was to compare my own familiarity with idol groups to their objective success and see whether my intuition matched what the data showed.

Next, I prepped the dataset for modeling by engineering the features I wanted to include. I converted debut dates into debut years so I could filter by K-pop generations and allow the model to pick up patterns such as whether third-generation groups tended to be more successful, and also created a company tier system. I put the major powerhouses like JYP, YG, SM, and HYBE/Bighit as the tier 3 companies. Then I put the mid-tier companies or HYBE subsidiaries that are somewhat known but not overwhelmingly influential as the tier 2 companies. Lastly, I put companies with very limited resources and/or little to no public presence as tier 1. After that, I then encoded group gender numerically to see whether boy groups or girl groups had any measurable differences in success.

Regarding the success model, to compare groups across eras fairly, I created a combined success score using three popularity metrics: organic YouTube views, physical album sales, and music show wins, and scaled each component so that no single variable would dominate the others. After examining the distribution of scores, I labeled any group with

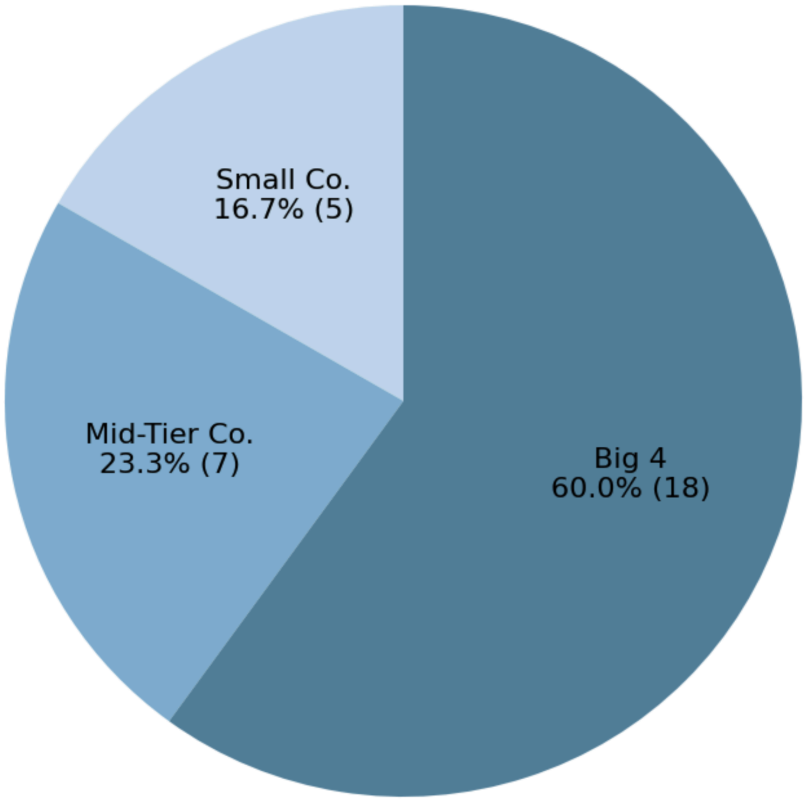
a score of 1500 or higher as “truly successful.” This cutoff was very exclusive, but it helped distinguish the groups that truly broke through and achieved immense global fame from those that stayed relatively unknown.

Since I had only one dataset, I applied an 80/20 training and testing split. I tried several modeling approaches that we learned in class, including multiple versions of random forest and SVC models, one logistic regression model, and a gradient boosting classifier. After testing them all, the random forest and gradient boosting models performed the best, each achieving an accuracy of approximately 79%, which was surprisingly high considering that I only used two features in the end, as I’ll expand on in the discussion section.

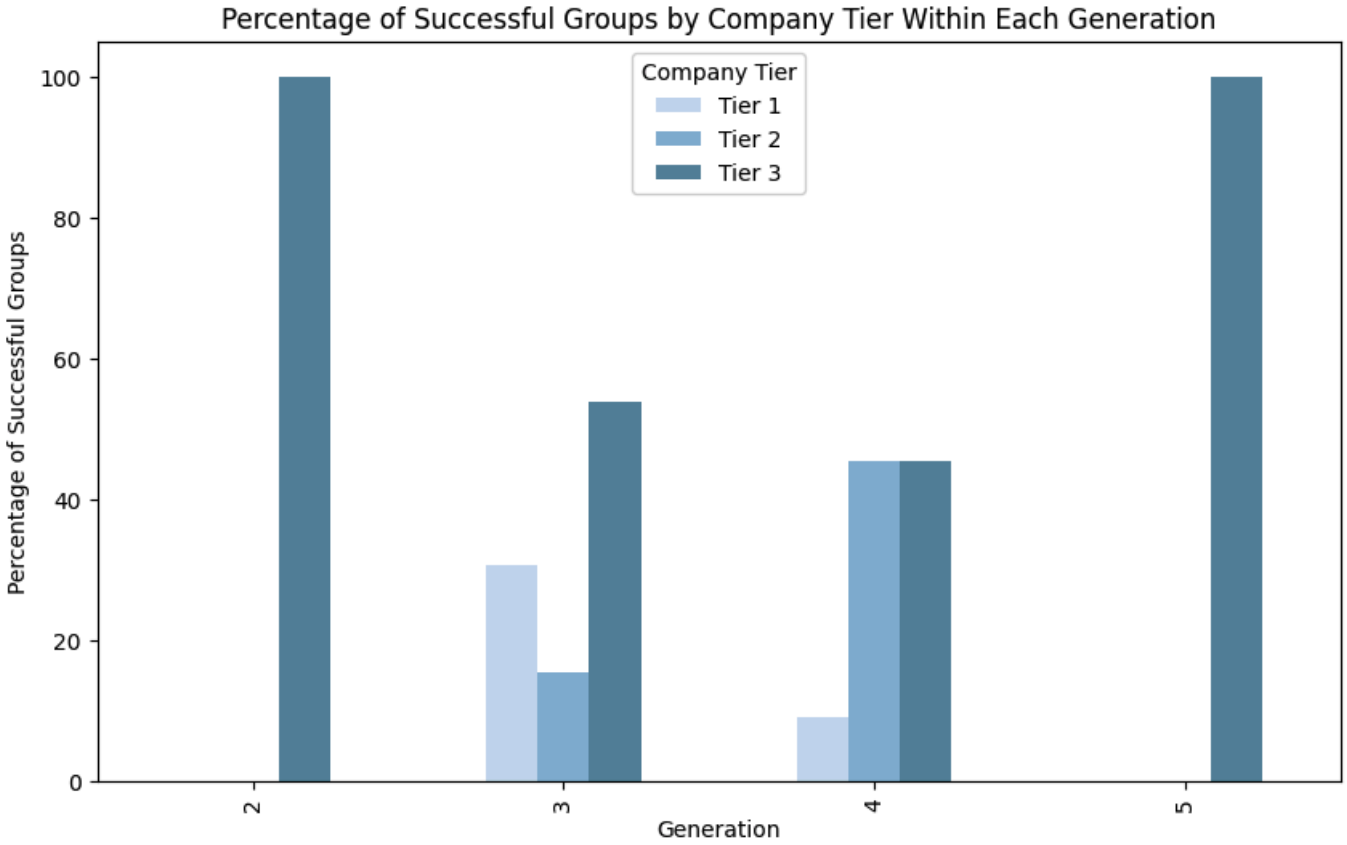
Aware of how metrics change for each generation, I decided to also examine each K-pop generation separately. Since I was limited on data and features, I decided to just tackle how much company prestige mattered through each generation, allowing me to see how the influence of big companies grew or shifted over time. Finally, for fun and insight, I compared my own knowledge scores against the model’s success labels to see how many of my “guesses” were correct and how well my personal familiarity aligned with objectively successful groups.

Results

Because my dataset was limited, the only factor I could confidently analyze was company prestige, and it ended up being the strongest predictor of success by far. The pie chart on the right shows the amount/percentage of successful groups that each company tier produced. From the chart, it’s evident that the clear majority of all successful groups in the dataset came from Tier 3 companies, while smaller companies produced almost none.



The generation-by-generation bar chart below also shows the same pattern over time, with big companies dominating every era.



Even with only two features (company tier and group type), my best models still reached around 79% accuracy, which at least suggests that success in K-pop is definitely somewhat determined by the company a group debuts under. Other factors in the dataset did not show strong or consistent patterns, so company prestige was the clearest and most reliable predictor.

Discussion

Once I started analyzing the dataset, it became clear that my original question, identifying all the factors that determine a K-pop group's success, wasn't actually possible to answer with the data available. The dataset was extremely limited and only included a few measurable features, so the most important factors like marketing, pre-debut exposure, song quality, and social media virality were completely missing. Because of this, I had to shift my analysis toward the only feature I could actually evaluate: company prestige.

This creates several caveats for interpreting the results. Since the dataset only allowed me to analyze one structural factor, the findings should not be taken as a full explanation of K-pop success. My success score, while useful for comparison, also does not capture older groups fairly because YouTube metrics did not exist during earlier generations. The models performed well mostly because the company tier dominated the dataset, not because they captured the true complexity of the industry.

Regarding any surprises I might've had, nothing in the results ended up surprising me. Even before modeling, I already knew that company power plays the biggest role in a group's debut momentum, and I also expected my personal knowledge ratings to align poorly with the objective success data because the dataset excluded so many real-world influences. Working with such limited information made the gaps very obvious.

Conclusion

With the limits of the dataset, the clearest takeaway from this project is that company prestige is the strongest and most consistent predictor of debut success among the features available. Even though my original question couldn't be fully answered, the analysis still showed an unmistakable pattern: groups from rich and powerful agencies have a significantly higher chance of becoming successful, regardless of generation. While this doesn't capture the full complexity of the K-pop industry, it highlights how uneven the playing field is and how much a group's future is shaped before they even debut. With more complete data, future work could explore additional factors, but based on what was available, the role of company power was the main conclusion that could be supported.