K-pop Insights

Understanding K-pop with Data Science.

02 - K-pop Groups Segmentation

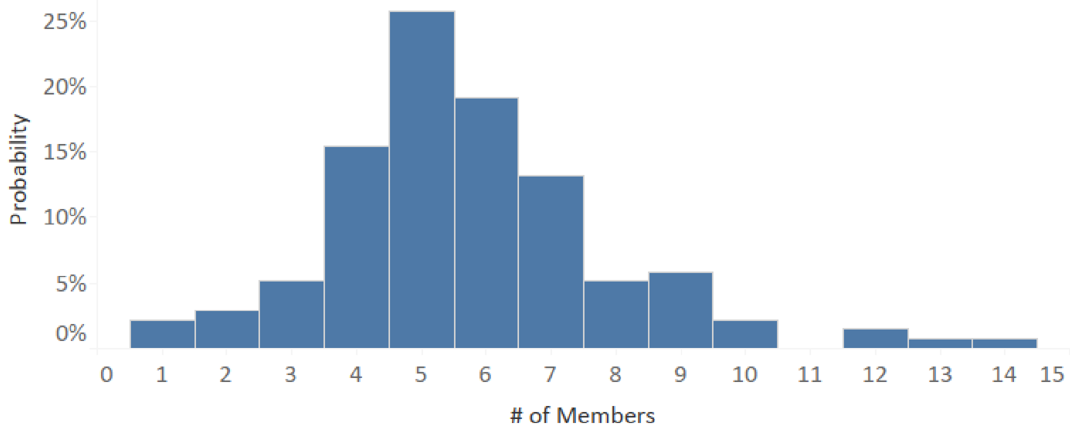
In this study, I used K-Means Clustering Algorithm to classify K-pop groups into 4 and found some interesting insights using the segmentation .

By Yuchen Wang.

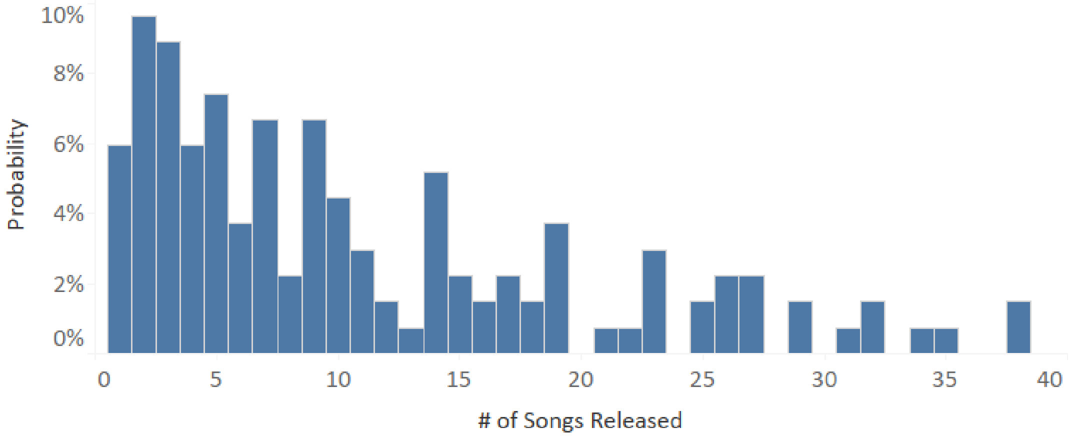
Data Source: 2005-2019 kpop data from dbkpop.com

Features Captured

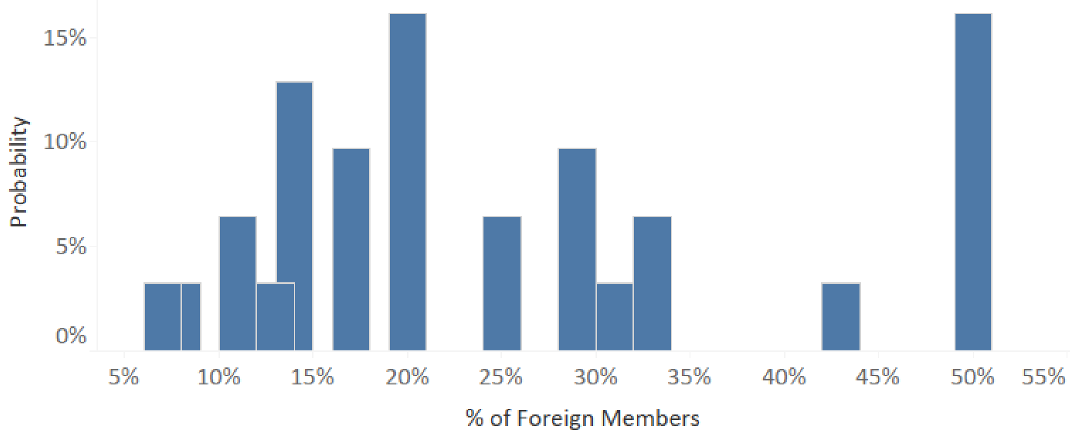
Group Size



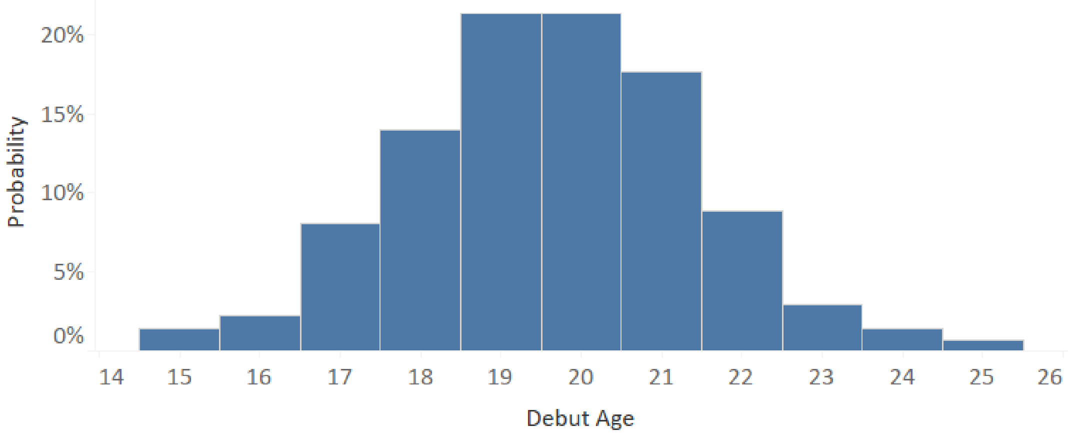
Production



International

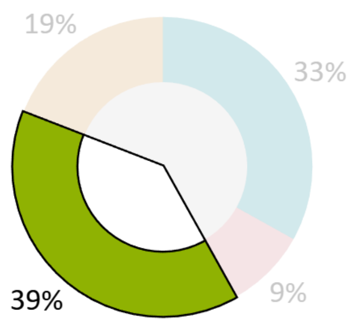


Youthness



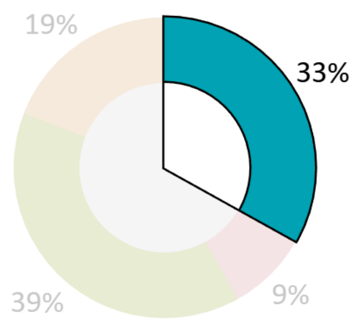
Clustering Output

Mature



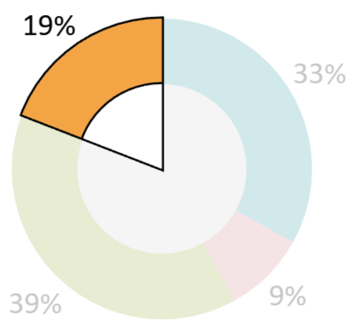
They debut at **21 years old** on average, 2 years later than the other groups, producing mature styles' images and songs .

Big-sized



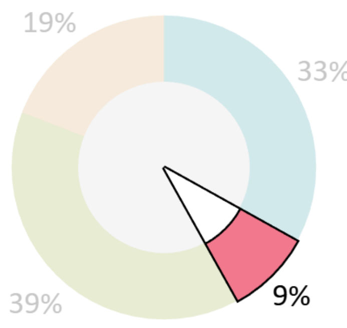
They averagely have **~8 members** in each group, bringing more variety of idols' images, personalities and skills .

Productive



They behave more productively than other groups, normally with **25+ songs** launched.

International



Rather than Korean local market, they focus more on international fans by recruiting more **foreign members (40% on avg.)**.

More in the Future

Some spoiler alerts:

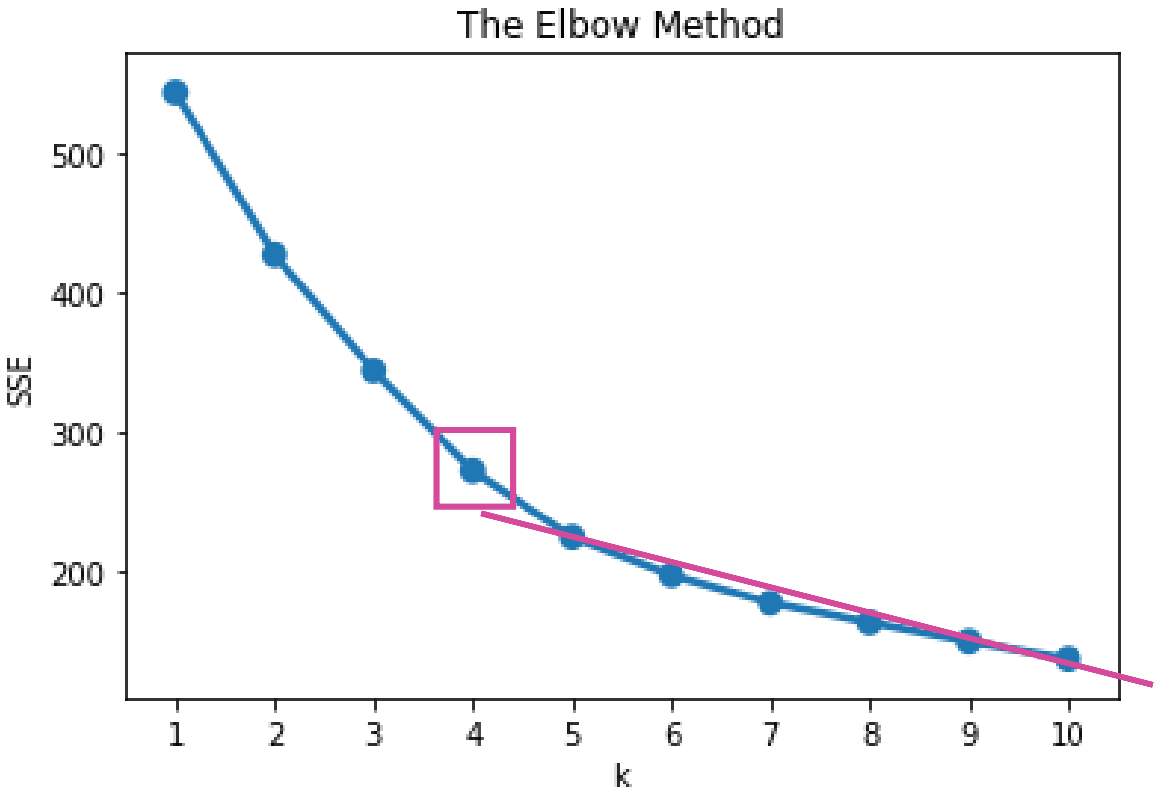
- Dashboards on **PowerBI** and **Google Data Studio**
- Music Videos Sentimental Analysis (**NLP**)
- Music Album Covers Analysis (**Image Processing**)

If you are interested in working on this fun project with me, please feel free to **contact me**!
linkedin.com/in/yuchenwang01

All codes, reports and dashboards are at my **Github**.
github.com/brantgithub/K-pop-Data-Aanalysis

Appendix – The Elbow Method

To tune the parameter of KMeans Clustering Algorithm, I used “The Elbow Method” to determine the number of clusters we want to output.

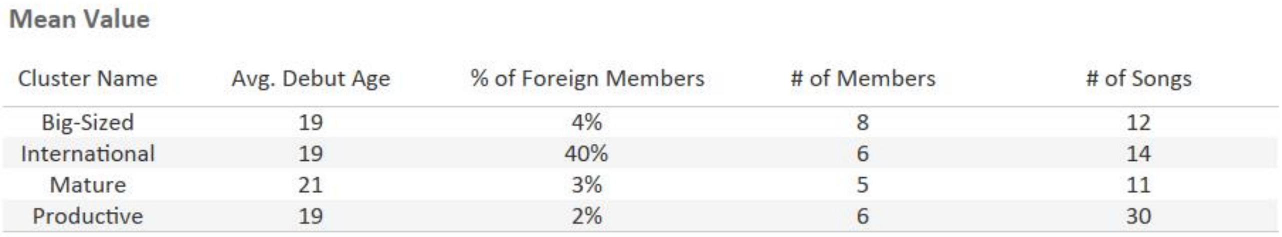


- X-axis: value of K
- Y-axis: SSE of the models

- To use this plot, we choose the K-value that will have a linear trend on the next consecutive K.

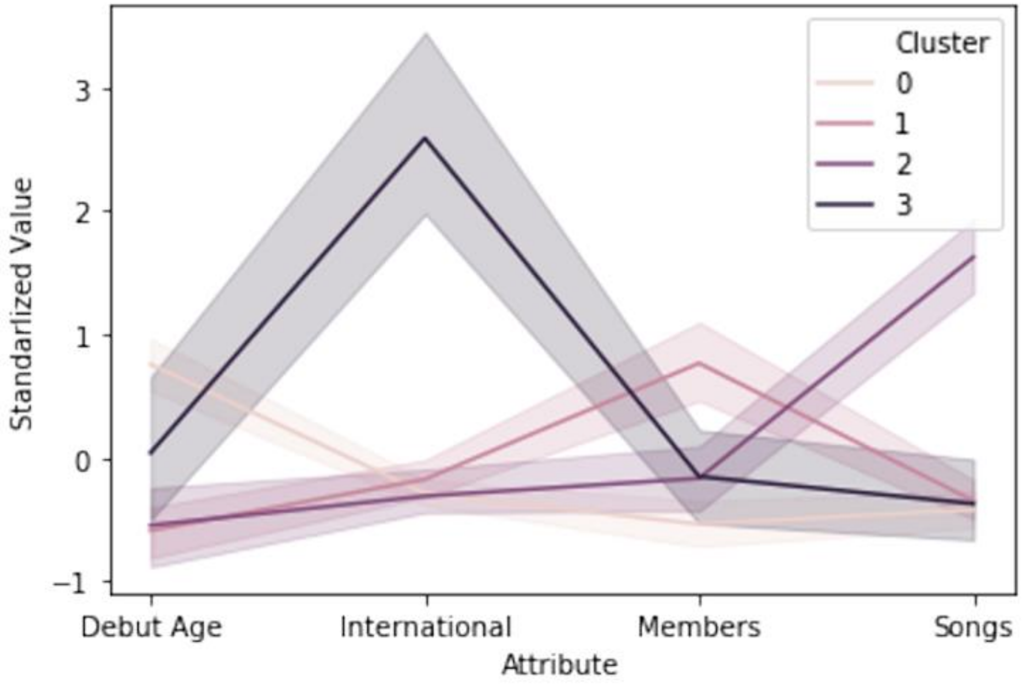
- Based on our observation, the K-value of 4 is the best hyperparameter for our model because the next K-value tend to have a linear trend.

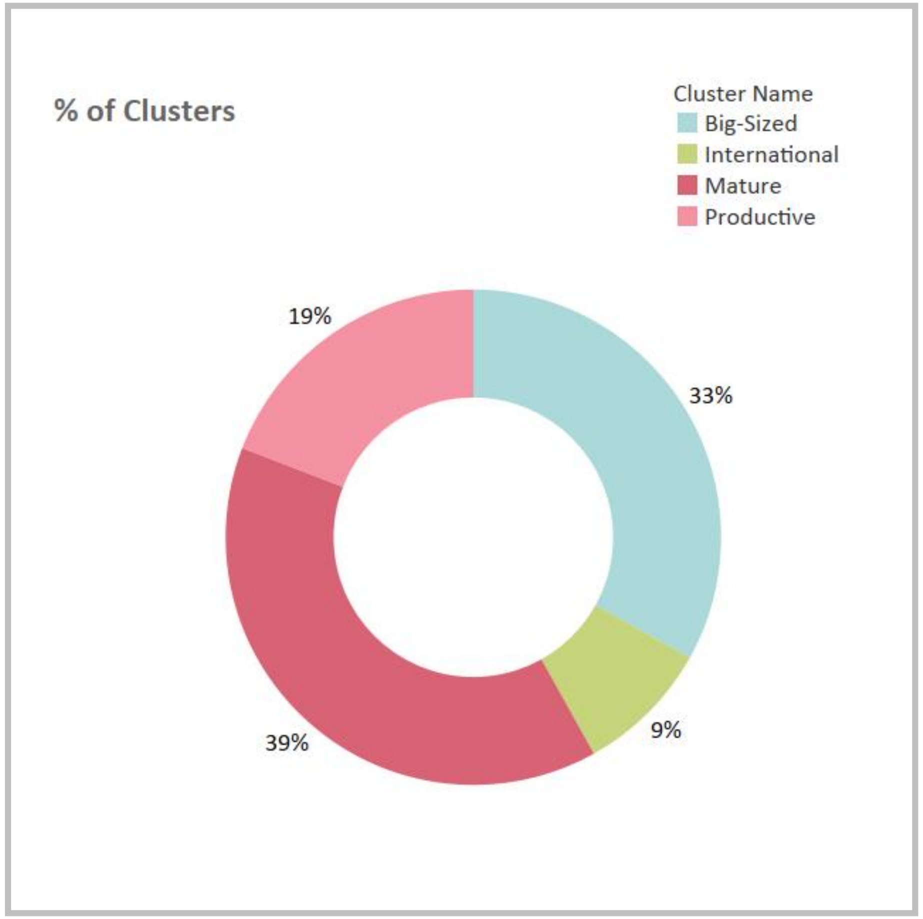
Appendix – Clustering Output



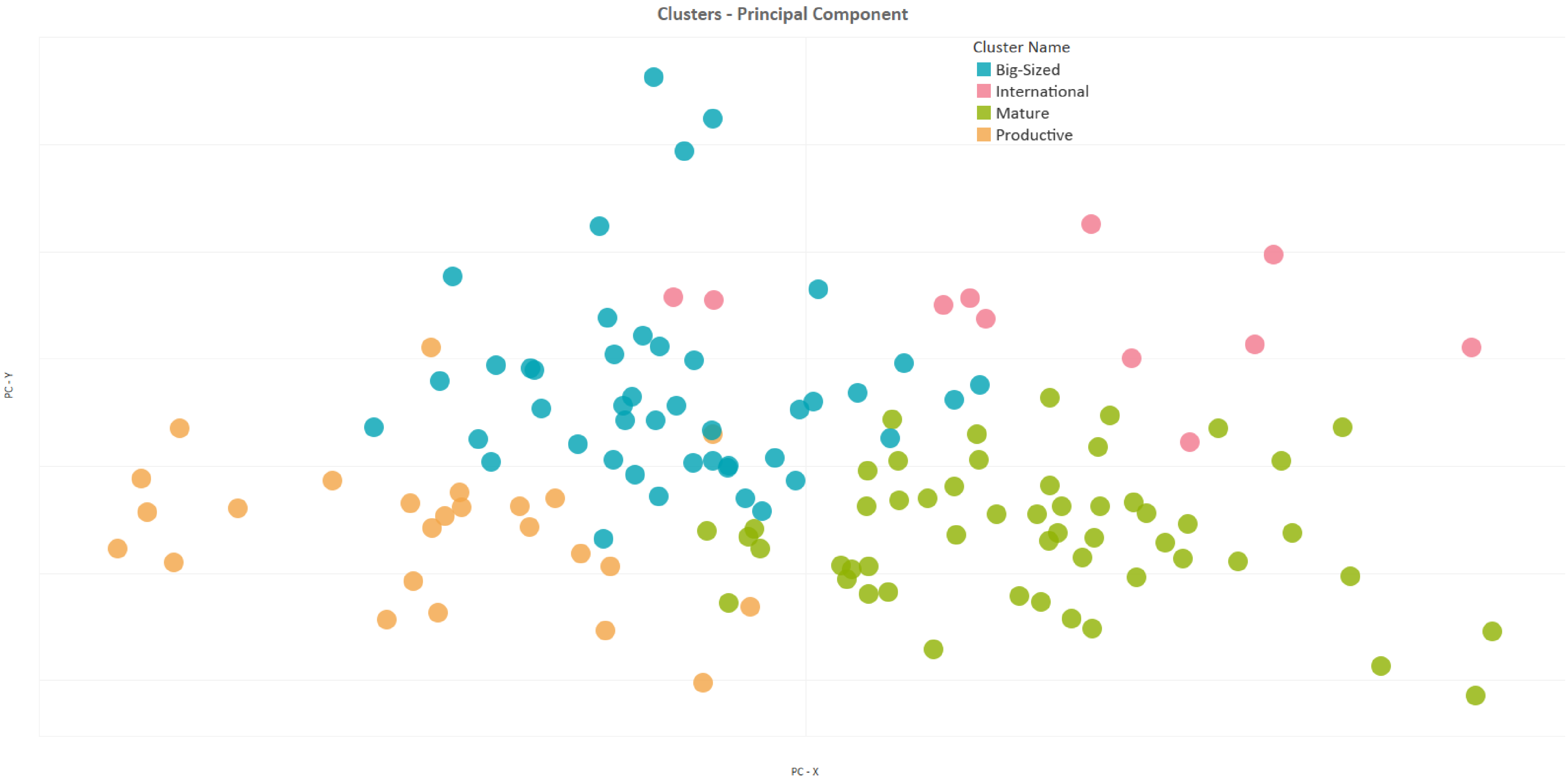
Mean Value

Cluster Name	Avg. Debut Age	% of Foreign Members	# of Members	# of Songs
Big-Sized	19	4%	8	12
International	19	40%	6	14
Mature	21	3%	5	11
Productive	19	2%	6	30




% of Clusters

Appendix – Principal Component Visualization



Clusters - Principal Component

Appendix

- Reference: https://towardsdatascience.com/customer-segmentation-in-python-9c15acf6f945
- Tableau Dashboards: https://public.tableau.com/profile/yuchen.brant.wang
- Reports, Codes & Models: https://github.com/brantgithub/K-pop-Data-Analysis