Data Nutrition

Health facts about our final cleaned dataset

Overview:

This dataset was created for our INFO 201 final group project, which aims to show how Korean media has spread to the US over time. Our final dataset has 55 months, totals of Netflix and Spotify Korean media appearances, a sum of both, 3-month and 6-month rolling averages, a rate of change, and a categorical value for the trend. Leading us to be able to make good predictions about the future, and analyze the past.

Data Creation Range: February 2017 - September 2021

Created By: INFO 201 Group: Austin, Ross, and Yashita

Content: Trend data

Sources: https://www.kaggle.com/datasets/dhruvildave/spotify-charts

https://www.kaggle.com/datasets/shivamb/netflix-shows

Dataset Nutrition Label the lower the alert, the better.

Alert Count	3
Completeness	1
Misrepresentation	0
Collection	1
Description	1
Composition	0

Consulting with our group, we have found that we have 2 main alerts in our dataset.

Completeness: Our dataset doesn't have as much data going as far back as we wanted. We are not able to safely and properly identify the start of the international rise of Korean media, such as Gangnam Style in 2012, as our data only goes back to 2017. We don't think this is a huge problem though, as it shouldn't interfere too much with our current project goal.

Collection: We couldn't fully verify the data that we collected, while we did our best, by making sure that we picked popular datasets on Kaggle, we cannot be 100% sure it is accurate as of now. Though we are highly confident due to the popularity and ratings of these datasets, and how they have been used in other analytical tests before.

Description: We don't have the best description for our dataset that we could have as of right now. We had to sort of change the topic of our project a few times, so our end goal isn't AS clear as it will be in the coming days, though we still have a great idea and plan to execute. Thus we are getting a point here because of this temporary issue.

8 Human Rights Principles in Data and Where We Stand:

Privacy: We ensured privacy by using aggregated data from public sources like Kaggle, without any personal identifiers.

Accountability: We take accountability for our data analysis, ensuring accuracy and responsible use of data in our research.

Safety and Security: Our data handling practices prioritize safety and security, preventing unauthorized access or misuse.

Transparency and Explainability: Our methodology and data sources, including Spotify and Netflix datasets from Kaggle, are fully transparent and explainable, as we have shown with background research and our nutrition label.

Fairness and Non-discrimination: We aimed for fairness by treating all data impartially, without discrimination based on origin or genre, and gathering this data from sources that exhibit the same qualities.

Human Control of Technology: Our project maintains human oversight over technological processes, ensuring ethical use of data analytics tools.

Professional Responsibility: We adhere to professional standards in data science, responsibly using data from Kaggle to inform our research, and not making any claims that we cannot backup with this data.

Promotion of Human Values: Our research promotes human values by exploring cultural exchange and its impacts on global media consumption patterns.

Our use case/questions:

1. Correlation between Korean Spotify consumption and Korean Netflix consumption
2. Are there specific songs that boosted Netflix Korean media production over others?
3. What platform shows the greatest increase in Korean media consumption over time?
4. Are we able to predict future growth with this modal?
5. What is the average monthly percentage change in content consumption for both platforms?
6. Growth consistency as well as summary statistics