Machine learning model for rating prediction based on Twitter engagement

Martin Veselinov Vladimirov 220116

DISCOVER YOUR WORLD

BUAS/AI&DS

Friday 20 January 2022

Breda University OF APPLIED SCIENCES

Abstract

As a leading media and entertainment company, Banijay is constantly seeking ways to improve the success of its television shows and better understand the preferences of its audience. One important factor in determining the success of a show is its viewer ratings,

This report presents a study on the use of machine learning techniques to predict ratings for Banijay Studios. The study utilizes a dataset of past ratings, shows, as well as twitter engagement, such as retweets, likes, quotes, and replies. The study employs machine learning techniques, specifically linear regression. The results of the study can provide valuable insights into the use of social media data as a predictor of television ratings and aid in the decision-making process for television studios.

Index

1Introduction 3

1.1Data Gathering, Cleaning and Preparation 3

1.1.1Gathering data 3

1.1.2Cleaning and preparation of Data 3

1.2Data analysis 4

1.3Twitter Data preparation and analysis 8

1.3.1Twitter Data analysis 10

1.4 Machine learning model 12

1.5 Ethics of study and Banijay Studios Benelux 13

2References 15

1 Introduction

The use of social media platforms, such as Twitter, has become an important tool for television studios to engage with audiences and promote their programming. In recent years, there has been a growing interest in using data from these platforms to predict the success of television shows. The present study aims to investigate the use of Twitter engagement as a predictor of ratings for Banijay Studios programming.

1.1 Data gathering, cleaning and preparation

1.1.1 Gathering Data

The data preparation process for this study involved collecting and pre-processing a dataset of past ratings and Twitter engagement metrics for Banijay Studios programming. The ratings data was provided by Banijay Studios Benelux. The Twitter engagement data was collected using the Twitter API, and included metrics such as tweets and retweets for each Banijay Studios show.

1.1.2 Cleaning and preparation of the Data

In order to ensure that the data was suitable for analysis, a number of pre-processing steps were applied. First, any missing or inconsistent data was removed. Next, the data was standardized to ensure that all variables were on the same scale. This step was important because some machine learning algorithms are sensitive to the scale of the input data.

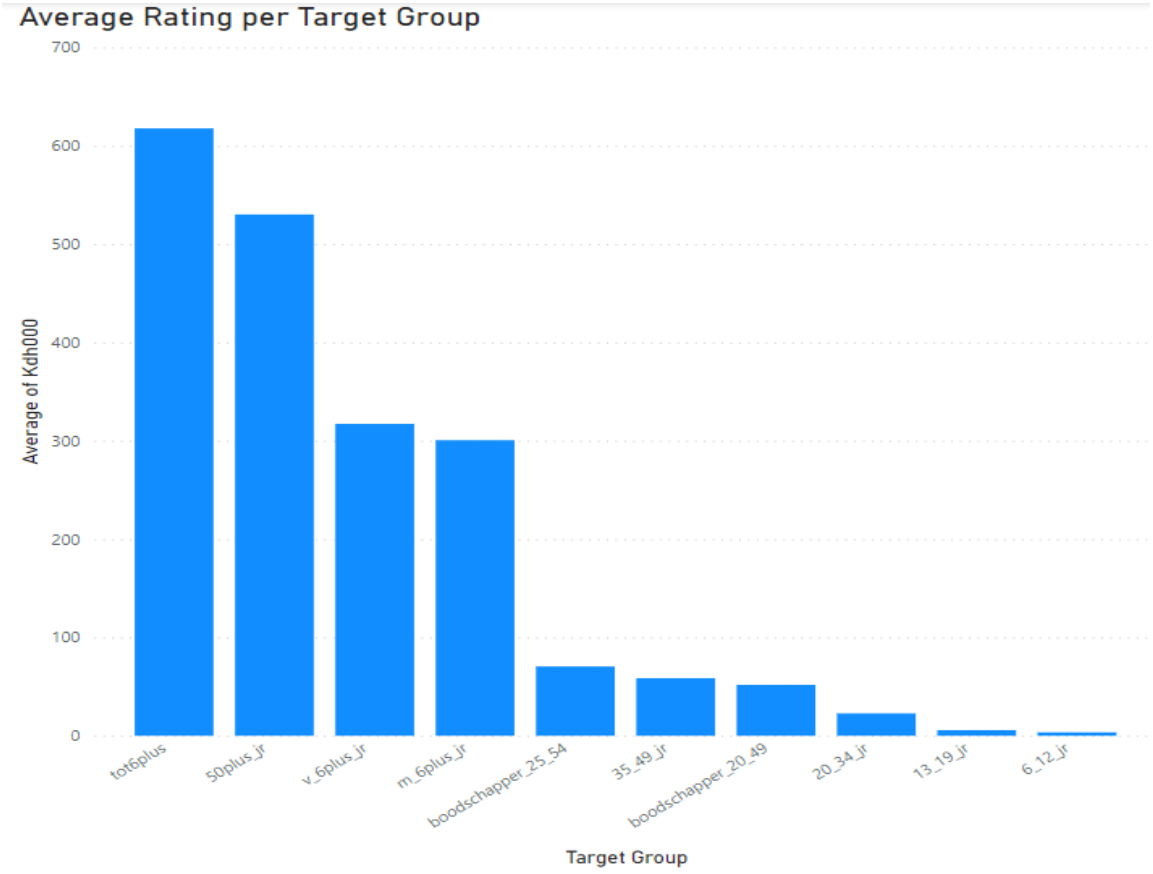
After cleaning the separate content and ratings datasets, merging was required in order to turn them into one flat file, so that analysis could be performed. Merging was done using the Python library “Pandas”. First, the library was imported alongside the two separate datasets. The next step was to check the datasets for any unnecessary information that they may contain. After removing the unnecessary and empty information, the datasets were sorted on datetime columns. Then, an empty variable for the data was created, as well as an index on which to merge the datasets. The column names were then retrieved from each dataset. After converting the necessary columns into the correct format, a for loop was made and based on indexing of the rows of the datasets, the two datasets were merged. After they were merged, the new data was combined with the combined

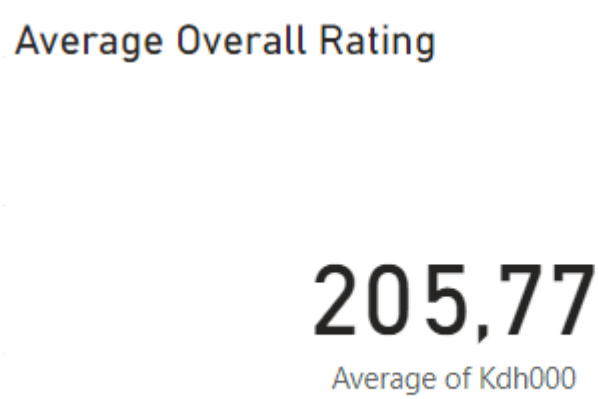
column names in order to put the new data into a data frame, which was then exported to csv files, one in full size, one compressed.

1.2 Data analysis

Data analysis allows organizations to identify trends and patterns in their data that might not be immediately obvious. This can help organizations identify new opportunities and potential problems before they become critical. By analysing data, organizations can make decisions that are based on facts and evidence, rather than assumptions or intuition. This can lead to more effective and efficient decision-making. Data analysis can help organizations improve their operations by identifying inefficiencies and areas for improvement. For example, it can be used to optimize production processes, improve supply chain management or identify customer needs. It can also aid organizations in identifying potential risks before they become critical as well as identifying a target audience. This can aid in making more effective marketing campaigns and boost the return on investment.

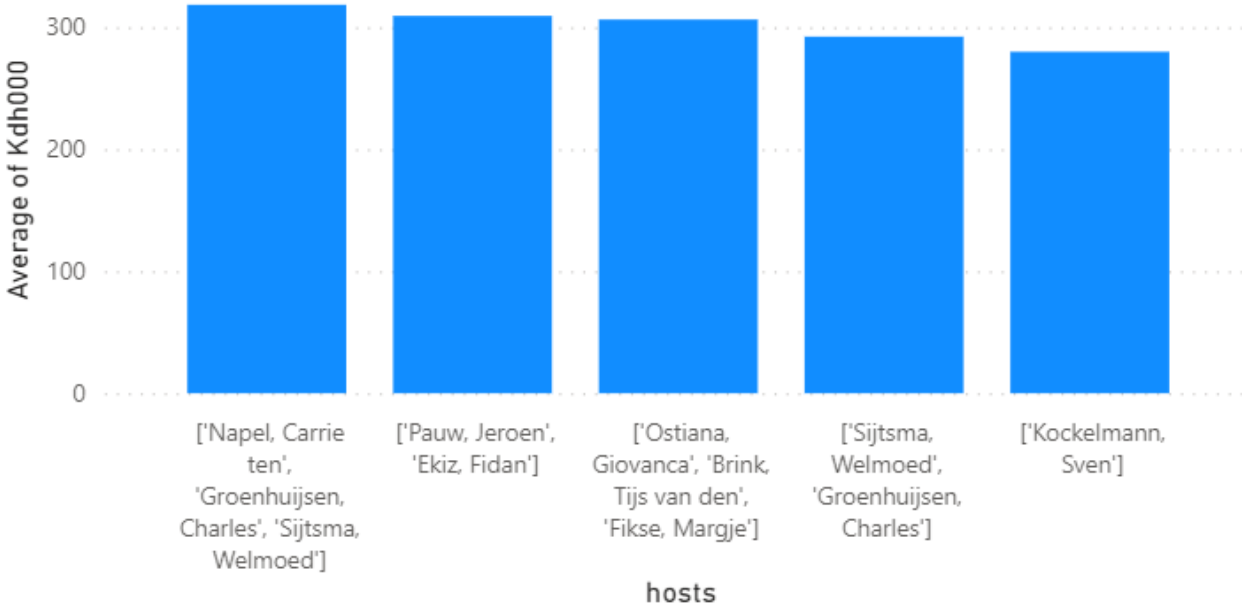
After merging the data into a flat file, analysis was performed on it. The analysis was done in PowerBI with visualizations in order to display the information in a visually pleasing and easy to digest way. First, a target audience analysis was done, to check the average rating per target group as well as the overall average rating across all groups.



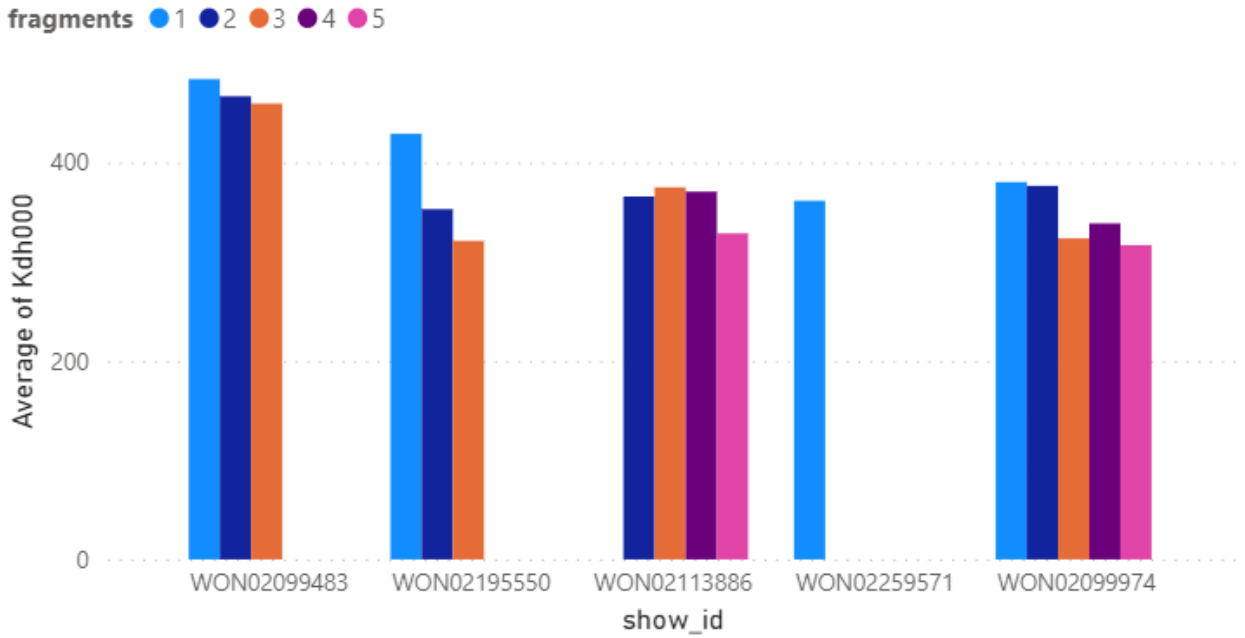


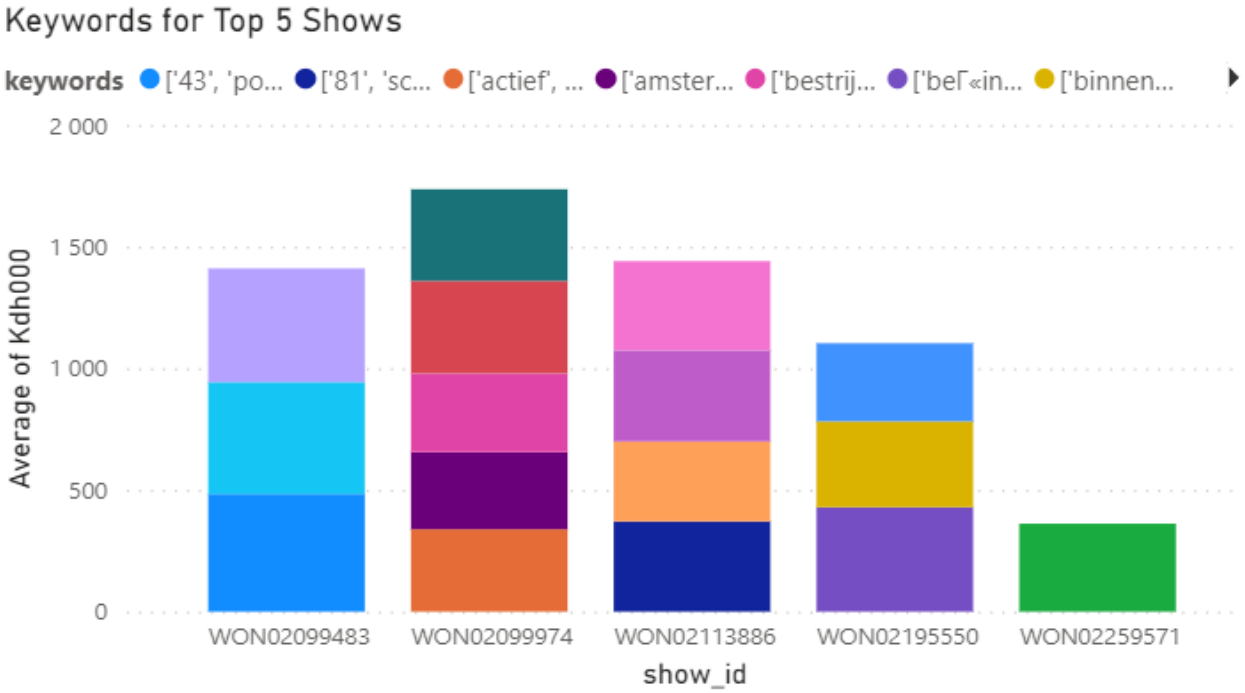
After this, an analysis was made on the content data. Visualizations were made on the top 5 hosts based on ratings, top 5 rated shows and their fragments, the keywords for the top 5 shows as well as a word cloud for the keywords.

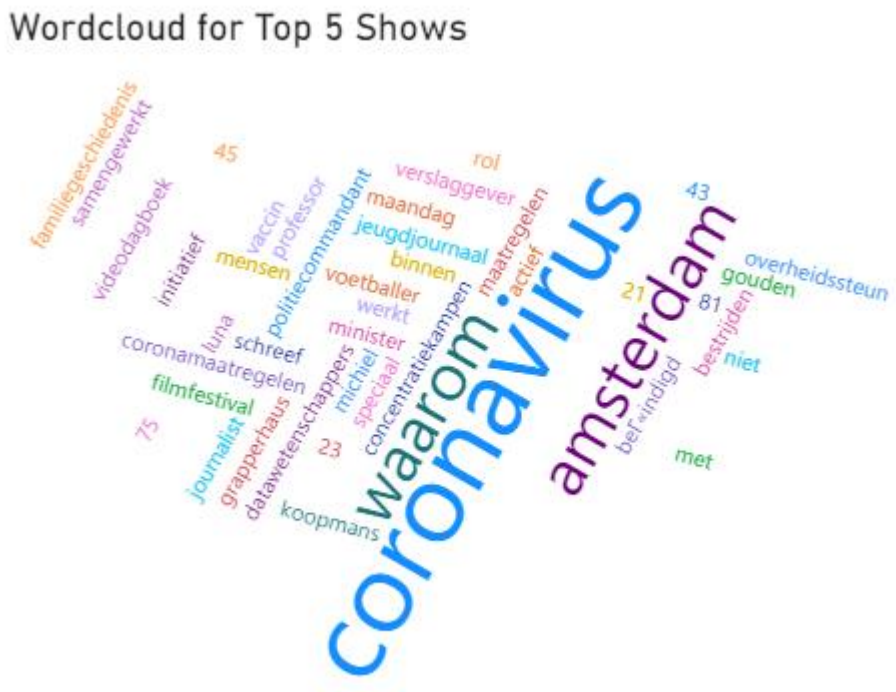
Top 5 hosts by Rating



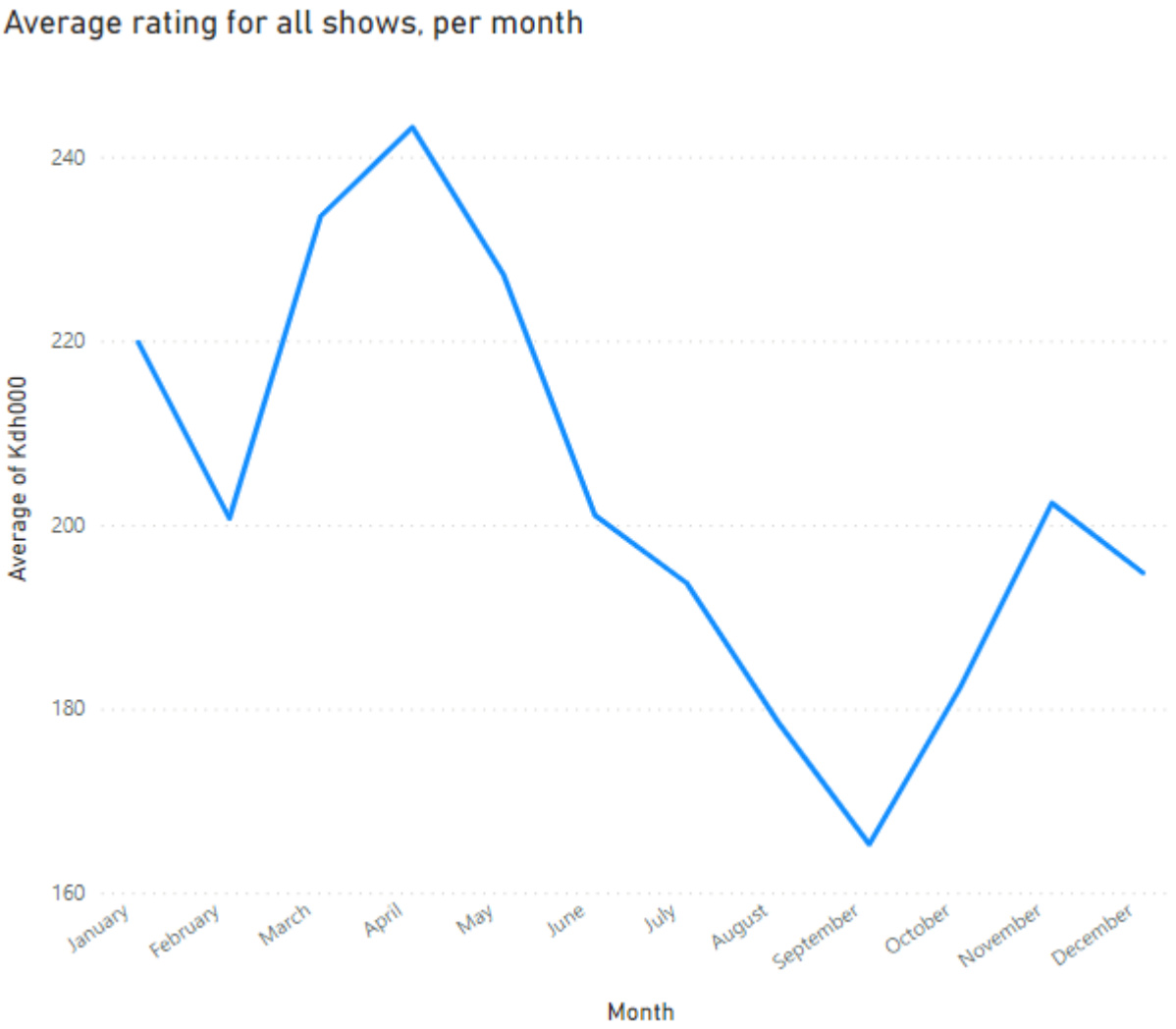
Top 5 Rated Shows and Fragments





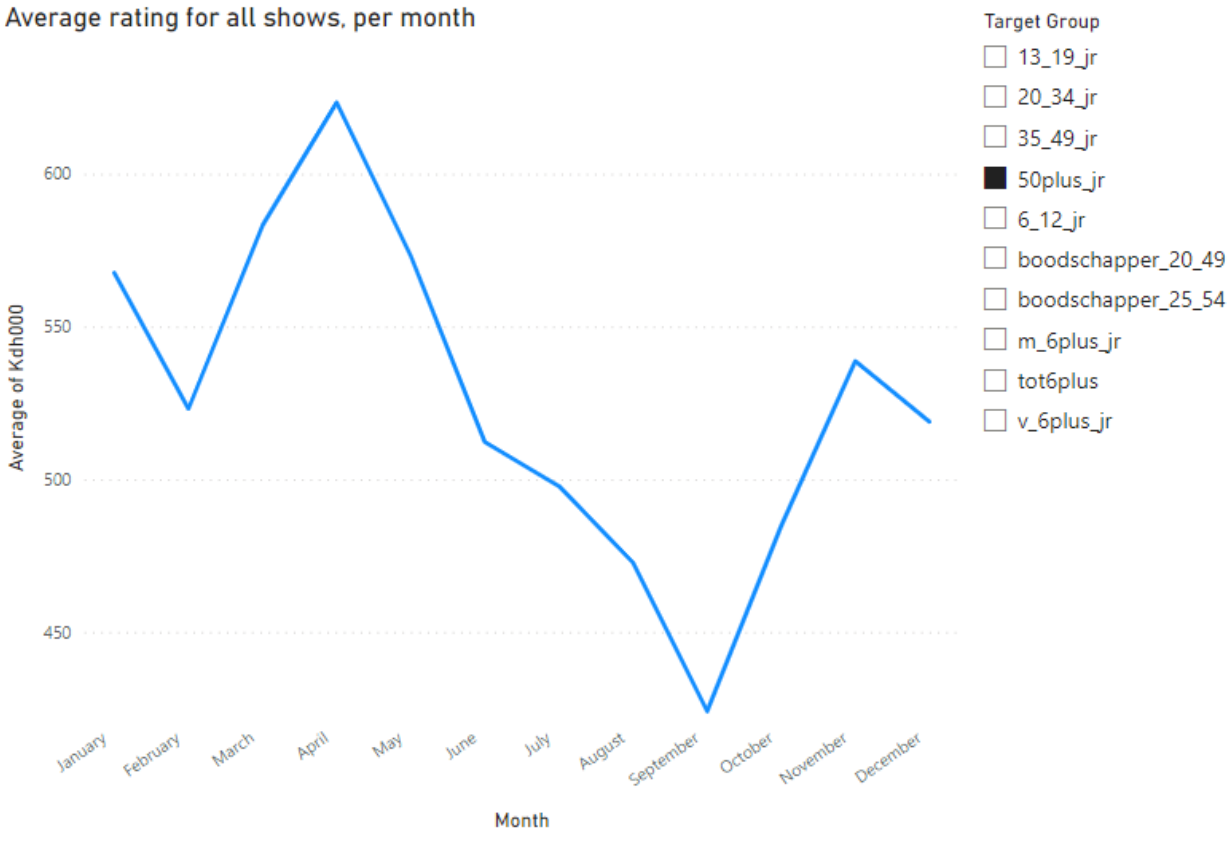


Lastly an analysis was made on the average rating per month. The analysis is represented by a line graph, which follows the month by month ratings. A slicer was also added in order to see the monthly ratings per target group.



It is noticeable that the ratings spike during the middle of spring, which is followed by a gradual slope, hitting its lowest point between September and October. This is then followed by a rise again, heading into the winter months.

The top highest average rating group is “50plus_jr”, which is to say, audiences over fifty years old.



Seeing the two graphs, it is visible that the ratings are mostly dependent on the 50plus_jr target group.

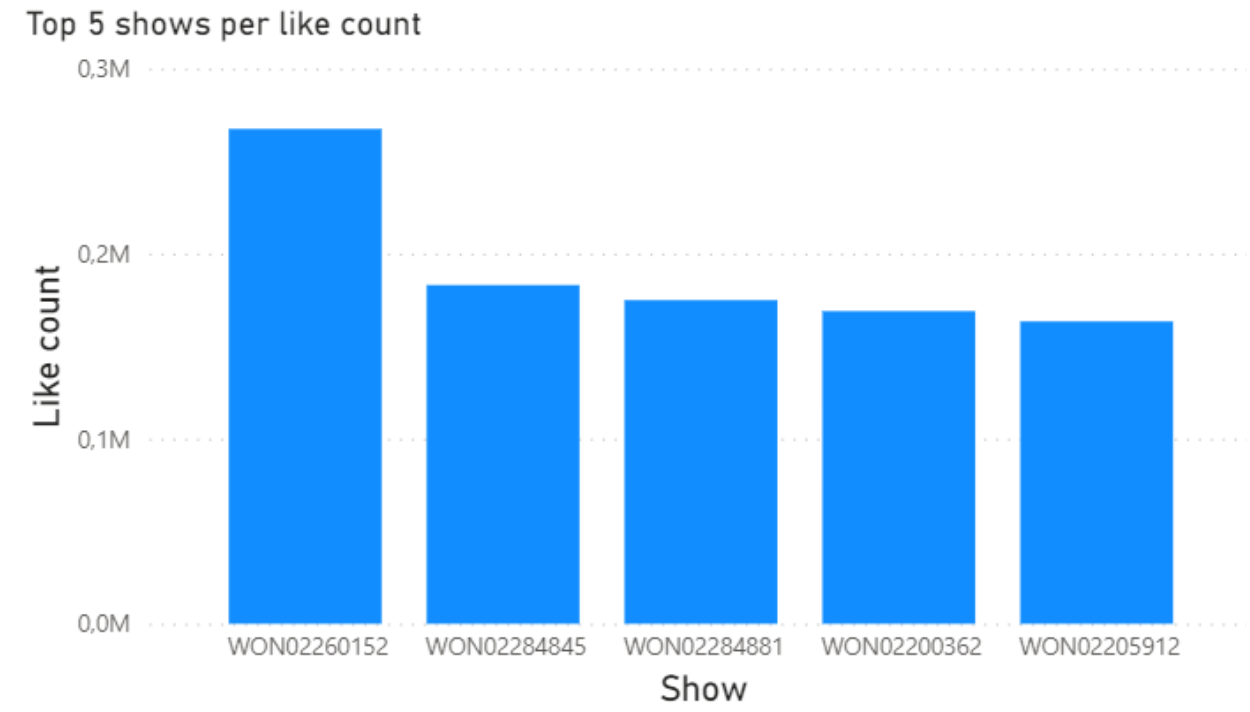
1.3 Twitter data preparation and analysis

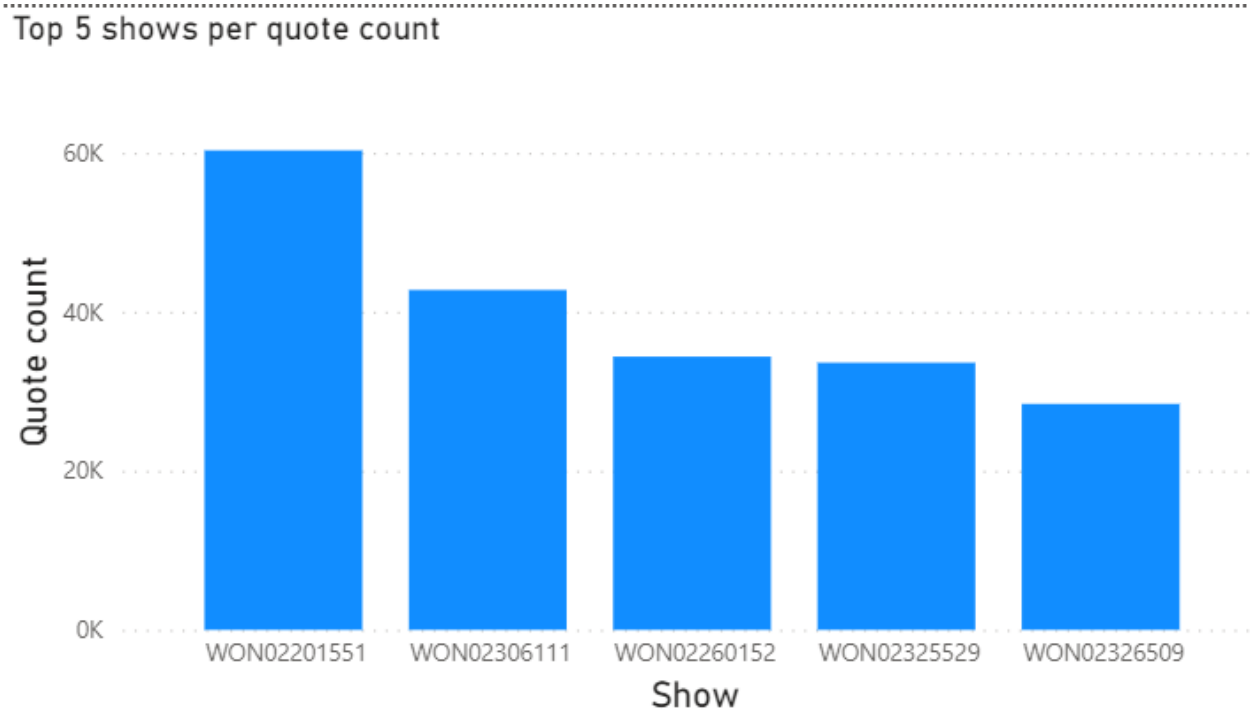
After the content and ratings data analysis concluded, the next step was gathering twitter data. In order to scrape data from twitter, a twitter API is needed. A Twitter API, or application programming interface, is needed to access and retrieve data from the Twitter platform. The Twitter API allows developers to access various types of data from Twitter, such as tweets, user profiles, and trending topics. It also allows developers to perform actions on Twitter, such as post tweets and send direct messages. The Twitter API allows developers to collect large amounts of data from Twitter, such as tweets, retweets, likes and user information, which can be used for various purposes such as sentiment analysis, trend analysis, and identifying key influencers. It can also be used to support research in various fields such as social science, marketing, or political science by providing access to large datasets of tweets and other data.

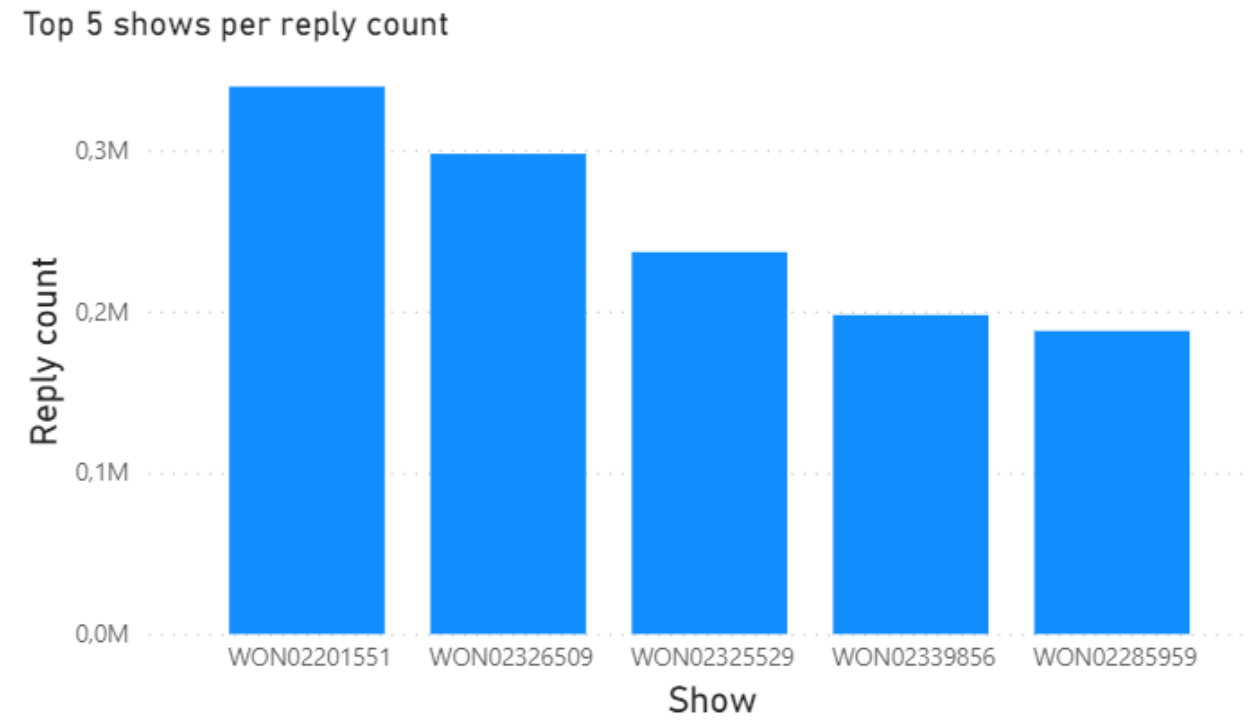
The twitter data has been provided by Banijay Studios Benelux. Firstly, the necessary Python libraries were imported. The libraries used were json and pandas. The merged flat file of the content and ratings data was imported, alongside the json twitter file. After that the twitter file was normalized in order for it to become a data frame for pandas to read. Converted the necessary columns to datetime and cleared the null values from referenced tweets. Created a new column “date” from the “created_at” column. Grouped the data by the date column on both datasets and merged the datasets on the date column.

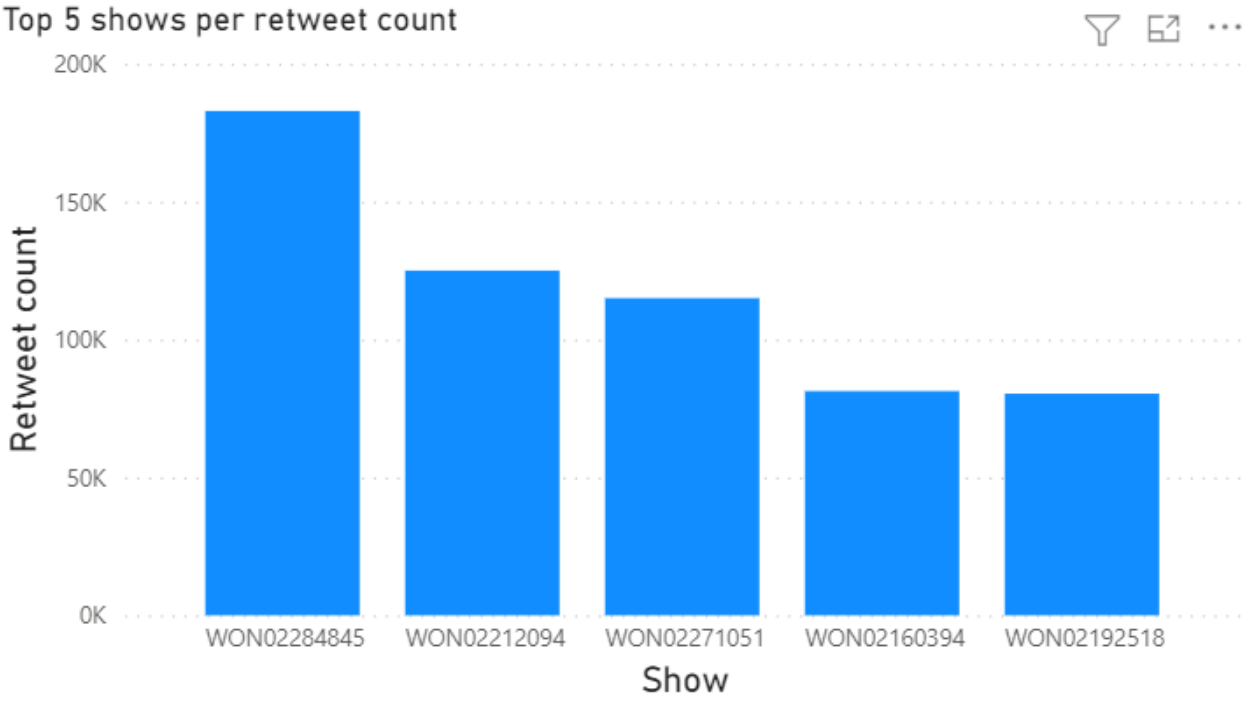
1.3.1 Twitter data analysis

After merging the Twitter data file and the Content and Ratings file, the new dataset was ready for analysis. The analysis consisted of checking the top 5 shows based on the Twitter engagement. This was done for each public metric separately.


Top 5 shows per like count
0,3M
0,2M
0,1M
0,0M
Like count
WON02260152
WON02284845
WON02284881
WON02200362
WON02205912
Show


Top 5 shows per quote count
60K
40K
20K
0K
Quote count
WON02201551
WON02306111
WON02260152
WON02325529
WON02326509
Show





In the above bar charts, we can see the top 5 shows per like, reply, quote and retweet count.

1.4 Machine Learning Model

This study's research question is as follows:" Could ratings for a show be predicted based on twitter engagement?". Using Scikit-learn's linear regression function this question was able to be tested. First, the Python libraries pandas, sklearn, sklearn.linear_model and sklearn.model_selection were imported. From these libraries, the functions "Linear regression" ,

“train_test_split” and GridSearchCV were used for the model. Then the merged file was read into the model. A parameter grid was formed in order to perform hyperparameter tuning later on. Afterwards, arrays were created in order to allow the model to make the correlation. There were two arrays which consisted of the following:
Array 1 – Like count, Retweet count, Quote count and Reply count.
Array 2 – KDH1000
Afterwards the data was split into training sets and test sets. The model was then created and fit. Then, the gridsearch object was created based on the model. After that the best parameters and the best score were printed out. Then a prediction is made on the test data with the best parameter. The last step was evaluating the model’s performance.

The results the model gave were that there is no correlation between twitter engagement and ratings based on the data provided. The resulting correlation was
R^2 = 0.00014707396339941337

Correlation refers to the relationship between two or more variables and how they change together. In statistics, correlation is a measure of the strength and direction of the linear relationship between two variables.

There are two main types of correlation: Pearson's correlation (also known as Pearson's product-moment correlation or PPMC) and Spearman's correlation (also known as Spearman's rank-order correlation). For the purposes of this study, Pearson’s correlation was used.

Pearson's correlation is a measure of the linear association between two continuous variables. It is a value between -1 and 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. Using the definition of this correlation, it is visible that there is no correlation between twitter engagement and a show’s ratings. Using machine learning for this business case, provides insight into the audience.

1.5 Ethics of study and Banijay Studios Benelux

The ethics section of this report is an important aspect of the research and it is essential to ensure that all data collection and analysis is conducted in an ethical manner. The following ethical considerations were taken into account during the course of this study:
Measures were taken to ensure that all data collected was kept confidential and secure. All data used in this study was publicly available on Twitter and was collected using the Twitter API. This means that the users who generated the data were not specifically asked for their consent to be included in the study. However, by using the data that is publicly available on the platform, it is assumed that the users have agreed to the platform's terms of service and privacy policy, which includes allowing the data to be used for research. The methods used in this study were fully disclosed in the report. This includes information on how the data was collected, how it was analysed, and how the results were obtained. This allows other researchers to replicate the study and assess its validity. The data used in this study was analysed in an unbiased manner and no attempts were made to manipulate the results to achieve a specific outcome. Additionally, the results of this study should not be used to discriminate against or stigmatize any individuals or groups. The researcher has a responsibility to ensure that the results of this study are used for the benefit of society and not for any harmful or malicious purposes. Additionally, the researcher must be aware of the potential consequences of their research and take steps to mitigate any negative effects.

Overall, this study has been conducted with the utmost attention to ethical considerations. The data collection and analysis followed the best practices to ensure the protection of data privacy and security, informed consent, transparency, fairness and responsibility.

Banijay is a company that emphasizes on creating a safe and respectful working environment for everyone, promoting equality and diversity both on and off-screen, providing adequate support to protect the well-being and dignity of its employees and contributors, and taking immediate action in case of any inappropriate behaviour. The company also prioritizes the physical and mental health of its staff and contributors, is committed to an open culture, has a goal of carbon-neutral production and is willing to evolve and reassess its welfare pledges as necessary.

On Banijay's website, a privacy notice can be found, which explains in full detail how Banijay Studios handles personal data. The purpose of the notice is to inform employees of Banijay (the "Company") about the collection and use of their personal information in accordance with the General Data Protection Regulation (GDPR). The notice explains that Banijay UK is the "data controller" responsible for deciding how the company holds and uses personal information, and that employees should address any data concerns to the Human Resources Department. The notice also lists the types of information that the company may collect and process, such as name, address, contact details, qualifications, skills, experience, employment history, remuneration, bank account information, national insurance number, marital status, next of kin, dependants, emergency contacts, nationality, criminal record, schedule, attendance, leave taken, disciplinary and grievance procedures, performance assessments, medical or health conditions, and equal opportunities monitoring information. The notice also explains how the company collects and stores this information. It explains the qualified rights of the employees of Banijay. Examples of this include:
access and obtain a copy of your data on request;
require the Company to change incorrect or incomplete data;
require the Company to delete or stop processing your data, for example where the data is no longer necessary for the purposes of processing;
object to the processing of your data where the Company is relying on its legitimate interests as the legal ground for processing; and
ask the Company to stop processing data for a period if data is inaccurate or there is a dispute about whether or not your interests override the Company's legitimate grounds for processing data.

2 References

Life at Banijay. (2022, 16 december). Banijay Group - We are Banijay. https://www.banijay.com/life-at-banijay/

Privacy Notice. (2022, 20 juli). Banijay Group - We are Banijay. https://www.banijay.com/privacy-notice/

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility

Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

PHONE
+31 76 533 22 03
E-MAIL
communications@buas.nl
WEBSITE
www.BUas.nl

DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES