Machine Learning to predict viewership

Mátyás Tatár
223073

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

Breda University OF APPLIED SCIENCES

Machine Learning to Predict Viewership: Is it Ethical?

Mátyás Tatár

Data Science and AI, Breda University of Applied Sciences

FAI1.P2-01 Project 1B ADS&AI 2022-23

Alican Noyan

January 20, 2023

Abstract

More and more companies are using big data to predict user behaviour. This report explores the journey from raw data to making predictions using machine learning, and the ethical concerns surrounding it. The objective was to use seemingly unrelated, separately documented data to build a machine learning model that would predict how much viewership an episode of a show would receive, if it aired with certain criteria. The results of the model are less than ground-breaking, but with more fine-tuning and exploration, they could have serious implications to the success rate of the company at which it is applied. An ethical prediction model is possible, it just requires careful navigation of morals, philosophies, cultures, and worldviews.

Index

1 Introduction

The Banijay Group has seen a rapid surge in popularity since they first started in 2008. In a matter of years, they were able to proudly call their own, countless top TV shows, like the Peaky Blinders, Black Mirror, and Survivor. Alongside these international co-productions, the Dutch branch of Banijay also produces a local talk show, OP1, which covers a wide array of trending topics such as politics and world events. As a result, they have asked for a machine learning model that predicts viewership based off of different data from the talk show. This is meant to help them understand what drives viewership, and ultimately, increase their success rate for future episodes, or even different shows. Building such a model requires access to large amounts of viewer data, and in turn leads to ethical implications which must be dealt with cautiously.

1.1 Introduction to datasets

Luckily, the Banijay Group provided curated datasets that are free of privacy-sensitive, and personal data for OP1. The first dataset provided is a “content” dataset consisting of vital information about the show, such as the date it aired, and its starting, and ending times, as well as the show id, title, keywords used, hosts, and many more. Next, we received a “ratings” dataset which contains viewership information such as the date of viewership, the rating type, target group, viewership, and the ratios of target group reached. Lastly, Banijay provided a twitter metrics dataset which contains public metrics of their own account. Data such as like count, reply count, author id, follower count, date, and referenced tweet (replies) can be found in this dataset. More privacy sensitive data can only be accessed directly by Banijay, and it is up to their discretion on how to deal with it, and whether to include it in future prediction algorithms or not. One example of privacy sensitive data may be the “referenced_tweets” column, since it mentions the user that was replied to. Although the implications may be miniscule, it is best to confer with the company or with privacy-regulating legislature to decide how to approach its use.

1.1.1 Cleaning and Pre-processing

The first step to creating a machine learning algorithm is having a single correlational dataset which contains values that are easily identifiable as belonging to a specific observation. This requires merging of the provided datasets. To start, both “content” and “ratings” datasets were cleansed of duplicated, or missing data. Next, a common key needed to be identified for merging. In the “ratings” dataset, each row corresponds to a specific broadcasting of the show. This means that the dates will be similar, if not exactly the same. To compare the dates of the two datasets, homogenous columns need to be created in both sets. This means a “date_time_start” and “date_time_end” column for the “contents”, and a “date_time” column for the ratings. These columns need to be formatted as a date time object to successfully, and accurately, compare their values. Furthermore, the show id provided in the “content” dataset needs to be split into its respective id, and segment (fragment) counterparts. This will also be used during the merging of the two datasets. Since the objective was to predict total ratings, the “ratings” dataset needs to be filtered to only show the “totaal” rating type.

To merge the “content” and “ratings” datasets, a lookup table is first created, in which the first column consists of the “date_time” from the “ratings” dataset, and the second column contains the matching show id from the “content” dataset. This is done by extracting the date time from the ratings and sorting it in ascending order. Then, a piece of code iterates through the dates in the lookup table and compares them to the dates in the sorted “content” dataset. If the code finds the lookup table to be in between the start, and end times of a broadcast, it extracts the show id from the currently iterated row and pastes it into the lookup table, next to the corresponding date. Once the lookup table is filled, it is merged to the “ratings” data on the corresponding dates, which is then merged with the “content” data on the basis of the show’s id.

The twitter metrics dataset was cleaned of tweets that were a reply to another tweet. This was done so that the dataset only contains tweets that may be of promotional value to a specific show. Next, the “created_at” column was transformed into a date time object to be able to compare date values

with the other datasets provided. To merge the twitter metrics to the content and ratings, the date part was extracted from the twitter set, and used as the merging junction.

1.1.2 Data Exploration

To explore and visualize the dataset, Microsoft Powerbi was used. First, the average viewership of Op1 was identified, which is seemingly low at 286.77. However, upon further analyses, the intended target group's viewership is extremely high (see Appendix 1). A clear preference of hosts can also be inferred from the data, as 'Kockelmann, Sven' and 'Muusse, Talitha', along with 'Pauw, Jeroen', and 'Ekiz, Fidan' are the hosts with the highest overall viewership (see Appendix 2). The popularity of segments can also be inferred from the data, which provides insight on how long viewers' attention span is, what content they are interested in, what makes them switch, and what keywords draw the most attention (see appendix 3). A cloud of keywords suggests that the more keywords are identified, the higher the viewership rating will be. Lastly, viewership trends can be identified when exploring ratings on a month-to-month, or day-by-day basis. Unsurprisingly, it follows the general schematic of, "if we have time, we watch tv". This can be seen in the drops of viewership in the summer months, especially in September (see Appendix 4), as well as February for some reason. On the weekly level, viewership tends to be higher in the middle, and the end of the week (see Appendix 5).

1.2 Machine Learning Model

Two machine learning models were attempted. The first idea was a decision tree classifier algorithm. However, this was discarded soon after, as it would require more formatting and processing of the data in order to make sensible predictions. It could have served as a way to identify what group a specific broadcasting of a show belonged to, but not to predict its viewership. In order to predict a continuous variable, a linear regression model is needed. The drawbacks of such models are creativity, and logic. These models are all classified as supervised learning algorithms. This means that the creator has complete control over the variables that are fed into the model to make a prediction, as well as the desired outcome through passing the model labeled results. This limits the accuracy of the model in what

variables are used to measure rating, because the possibility of two variables being correlated is very low. Finding the right features to predict viewership is the most taxing part of building a model. However, its advantages draw on the same reasons. An unsupervised model is more difficult to explain, and may also correlate variables that have no relation at all. A decision tree could very well serve as a robust prediction model if taken to a higher level of complexity, as it could account for multiple variables in predicting viewership. That, however, is for someone else to build.

1.2.1 Building the model

To be able to correctly build the model, the problem statement, "Can we predict viewership based off of keywords used, and the segment of the show?" was formed. Then, two variable x and y were created. X equals the keywords, and the segment of the show, and y is the viewership rating. These variables are then split into training and test sets and fitted to the model. Once fit, the model creates a straight line through the plotted data points, and uses the linear algorithm to make predictions.

1.2.2 Results

Using these variables, the model's effectiveness is questionable. Its accuracy score is not even one percent, and its Root Mean Squared Error is over 200, which means the model's predictions are, on average, 200 points off from the actual value. This may mean that viewership cannot be determined by the keywords used, despite the results of data exploration, and may not even correlate with the segment of the show. However, it may also be a lack of expertise regarding models and how their inputs work.

1.3 Is the Banijay Group an Ethical Organization?

1.3.1 Ethical Company

To determine a prediction algorithm as ethical, we must look at how the business that uses it operates as a whole. To start analyzing where the Banijay Group stands in terms of ethics, we ask ourselves whether or not they incorporate privacy policies, and if their behaviors towards employees and external

stakeholders are in line with a vision for a safe, and sustainable future. The Banijay group clearly adheres to widespread, and strictly enforced privacy regulations such as the GDPR, as outlined in their privacy statement. Upon further reading, the group also lists the several ways they collect data of employees and associates, how the data is stored, shared, and can be removed on request. In the sections detailing the specifics of data handling, they emphasize the coherence of their practices with rigorous data-handling legislature. They also demonstrate an appetite for diversity as an international company with headquarters in more than 20 countries, and a healthy distribution of staff from all areas of life.

1.3.2 Ethical Processes and Tools

The processes and tools created or used by an organization also reflects on its philosophy. So far, they have demonstrated strict adherence to ethical and privacy related regulations, and their website clearly boasts an international mindset. Furthermore, in their code of conduct, they outline responsibility, expectations, commitments, unity, and equality. They encourage free thinking, discourage limitations set by corporations, and strive for innovation through gathering like-minded intellectuals and creatives to provide unparalleled content for the masses. Even after only covering two of three pillars towards an ethical organization, a picture is being painted, where Banijay is a gleaming beacon of ethics.

1.3.3 Ethical People

The final aspect of an ethical organization is the behaviors of its professionals towards other parties such as suppliers, society, environments, and owners. These behaviors, along with their awareness of frameworks surrounding ethics and professionalism are the peak determinants of an organization's ethical capacity. Further outlined in Banijay's code of conduct is their pledge for sustainability, and drive for a carbon-neutral footprint. These ambitions, coupled with their enthusiasm and awareness for fair treatment and ethical processes create the environment for a high-performing, seemingly spotless organization.

1.3.4 Ethics Discussion

Seemingly spotless is used because despite taking the greatest measures to ensure ethical practices, one cannot ensure completely guarantee bias-free judgement in employment and operation. Whether that be the gender distribution of high-ranking staff, or the incidents that are kept behind closed doors. No organization can be truly ethical. One can only do their best to maintain their reputation and keep up their word to expel the feintest whisper of corruption in their midst.

One possible ethical concern comes to mind when thinking about the reasons for such a viewership predicting model. Will it be used to draw in viewers only to increase profits? Is so-called “click-baiting” ethical? In the future, Banijay could outline their use for the algorithm more clearly. How will they change their approach after learning what drives viewership?

1.4 Conclusion and Discussion

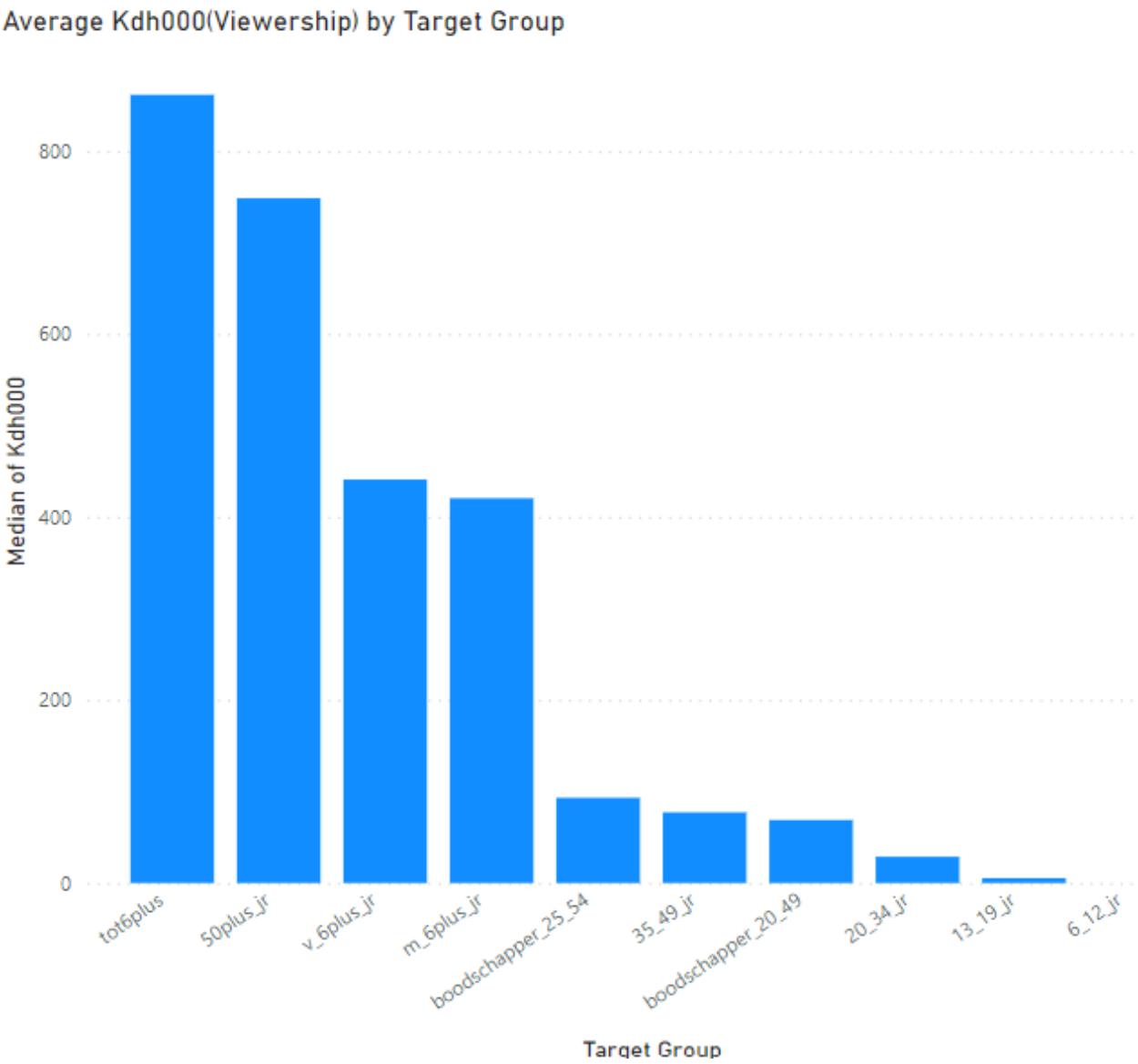
As a huge international media company that broadcasts a wide variety of shows, the Banijay Group has demonstrated their adeptness in navigating the ethical minefield that comes with dealing in big data. Their request to build a viewership predicting model came with a warning to follow ethical procedures from start to finish. We were clearly reminded that this data, even though stripped of identifiable information, is privacy sensitive, and should only be shown to parties that Banijay is associated with, or that are in accordance with GDPR regulations. Using data that contains privacy sensitive information may even be deprecated in the long run, as it leaves more room for error in building a robust model. Still, the resulting model of this exploration is wildly inaccurate, and has no real benefit to the organization. However, with more time and technical understanding, a model that has real benefit is definitely possible. That being said, it is also a possibility that there is just not enough data. The OP1 twitter account has a mere 330,000 followers. In the grand scheme of things, that is a very small dataset. Having so many followers does not mean 100% interaction from users. This severely limits the insights we can gain with the data, as it is not reflective of the actual target group, who might not even have twitter.

If Banijay continues to pursue a robust, transparent, and reliable rating prediction model, there is no doubt they will do so with the highest level of ethics in mind. They have even demonstrated consideration of IBM's three principles for AI in an organization. First, the model is meant to provide insights, and as a result, augment human intelligence. Second, they made it clear that this model is to be produced solely for their own use. Lastly, they made it clear through the requirement of this report that they seek transparency in the design and implementation of such a model.

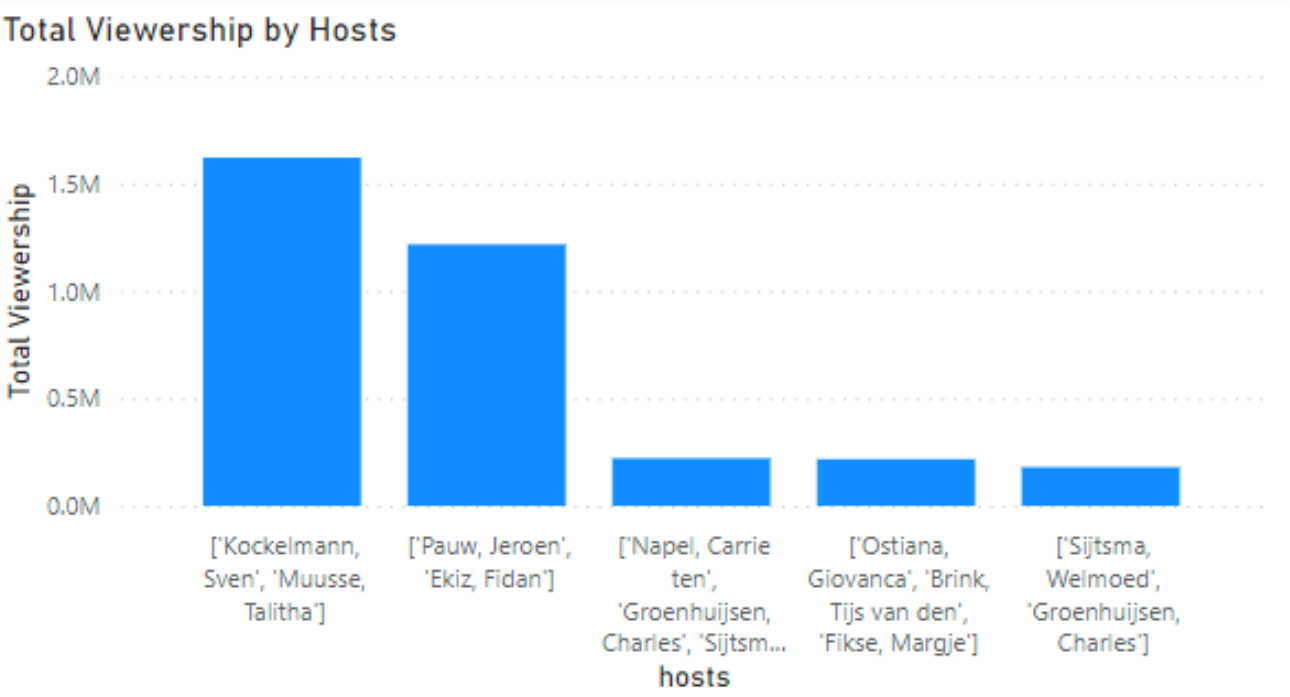
2 References

Appendix 1

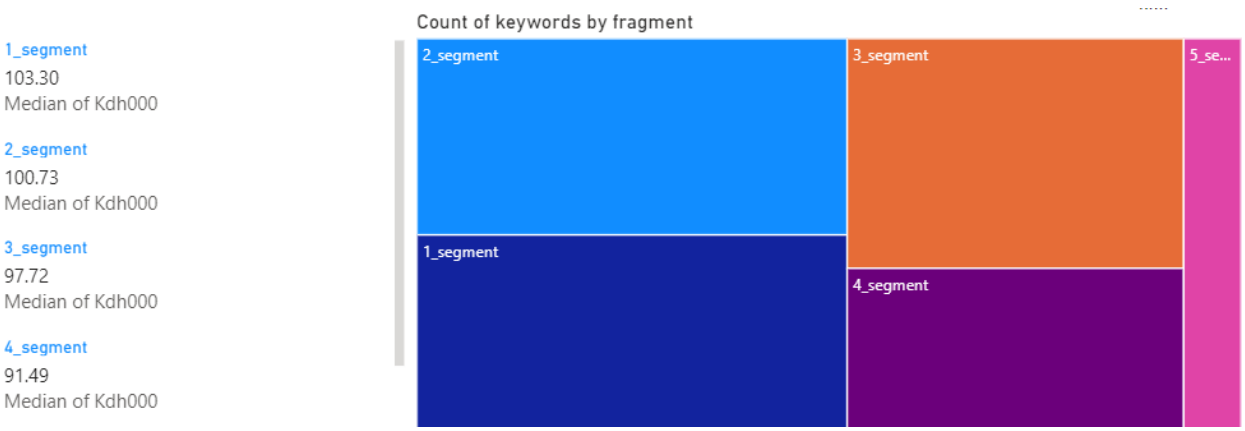
286.77	97.90
Average of Kdh000	Median of Kdh000



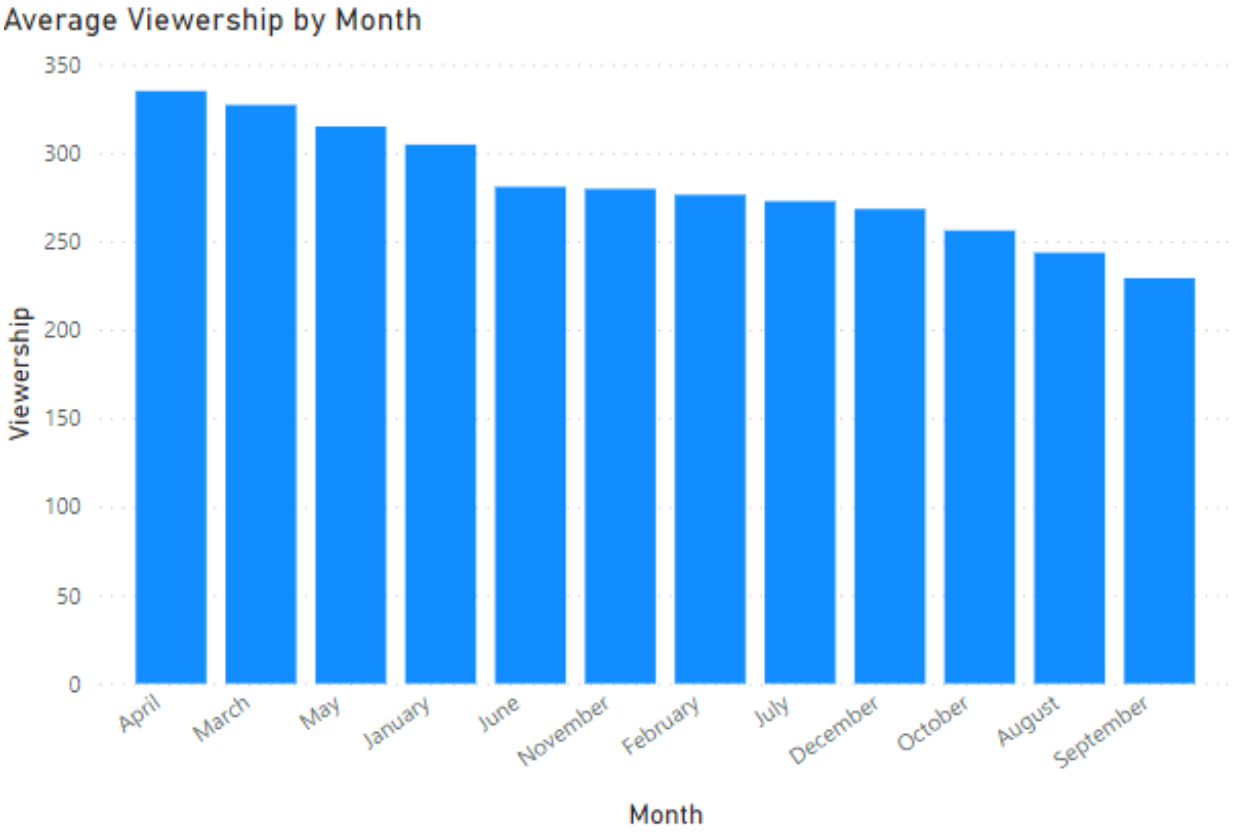
Appendix 2


Total Viewership by Hosts
2.0M
1.5M
1.0M
0.5M
0.0M
Total Viewership
['Kockelmann, Sven', 'Muusse, Talitha']
['Pauw, Jeroen', 'Ekiz, Fidan']
['Napel, Carrie ten', 'Groenhuijsen, Charles', 'Sijtsm...
['Ostiana, Giovanca', 'Brink, Tijs van den', 'Fikse, Margje']
['Sijtsma, Welmoed', 'Groenhuijsen, Charles']
hosts

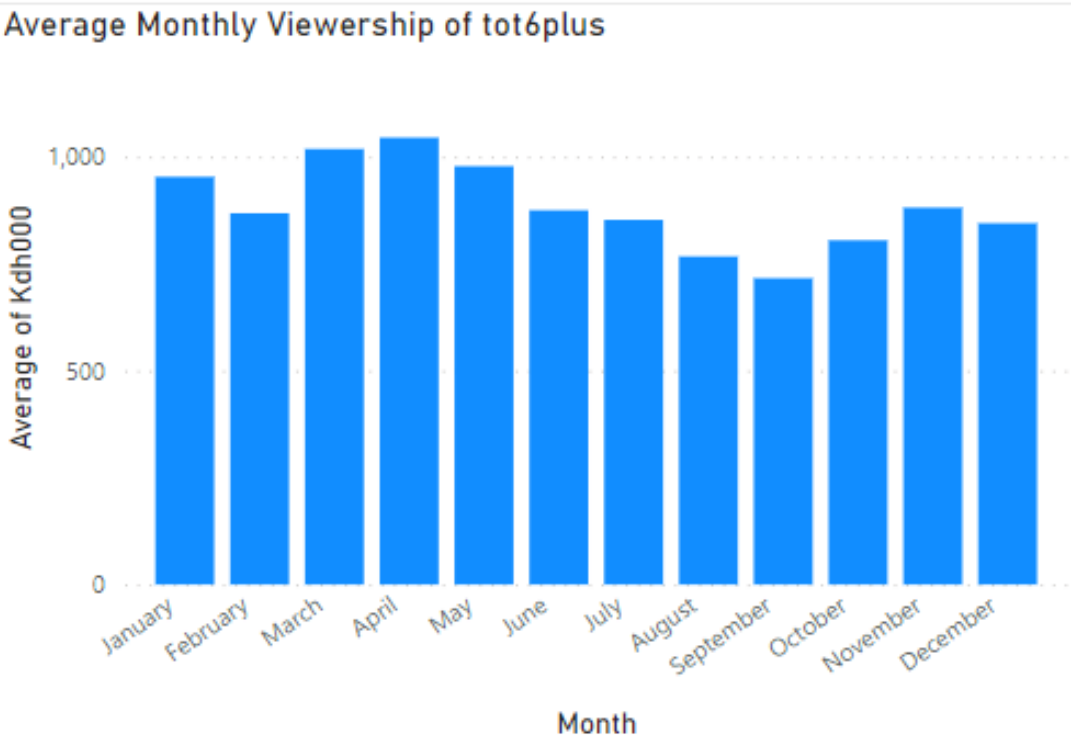
Appendix 3

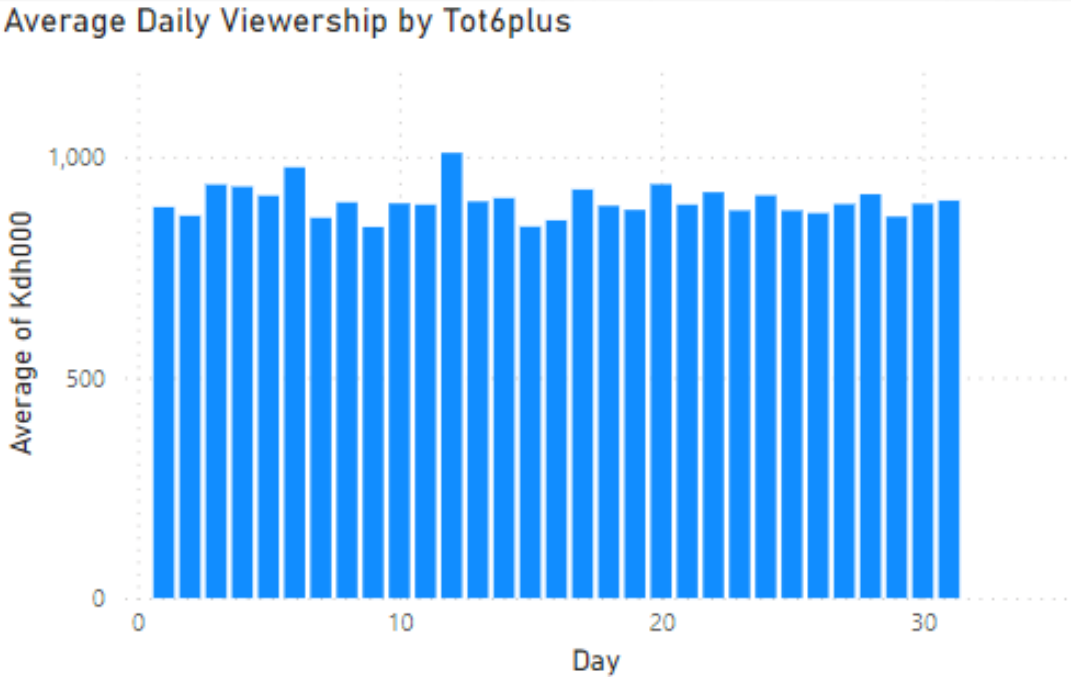


Appendix 4



Appendix 5


Average Monthly Viewership of tot6plus
1,000
500
0
Average of Kdh000
January
February
March
April
May
June
July
August
September
October
November
December
Month


Average Daily Viewership by Tot6plus
1,000
500
0
Average of Kdh000
0
10
20
30
Day

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