



VIDIO.COM

10% Data Driven Analytics

~ MUHAMMAD FADHIL ABIDIN ~

TABLE OF CONTENTS

Insight and Story	Exploratory Data Analytics			
Machine Learning Model	Random Forest vs XGBoost			
Tools	Tools that I used to work			
Platform To Watch	Answering question 4			
Top 10 Visitors	Answering question 5			



01

Insight and Story

Exploratory Data Analytics

Distribution of Average Bitrate of Play

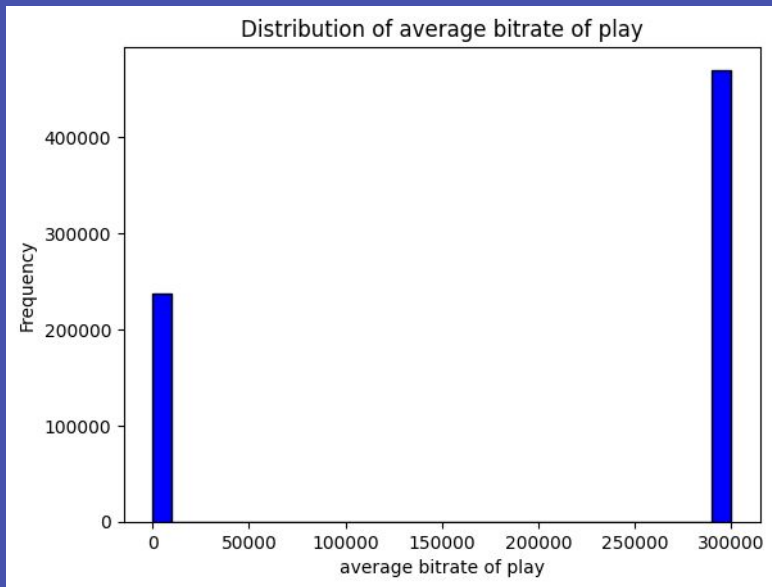
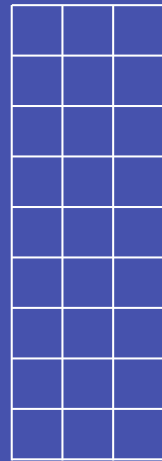


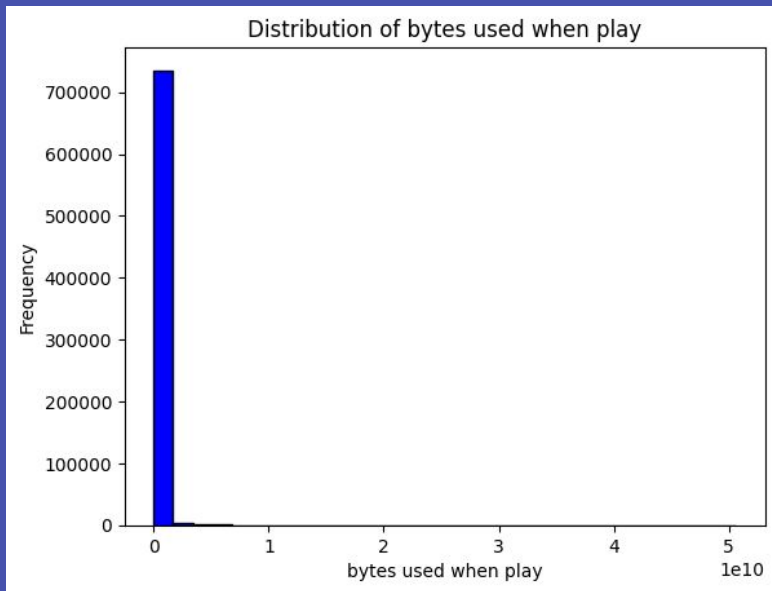
Diagram that prefer high quality tend to choose higher bitrate playback.

Diagram that are more concerned about data usage may choose lower bitrate playback.

Bitrate differences can also affect the ability of the device a user is using to play. Older or lower-spec devices may not be able to play high bitrate videos.



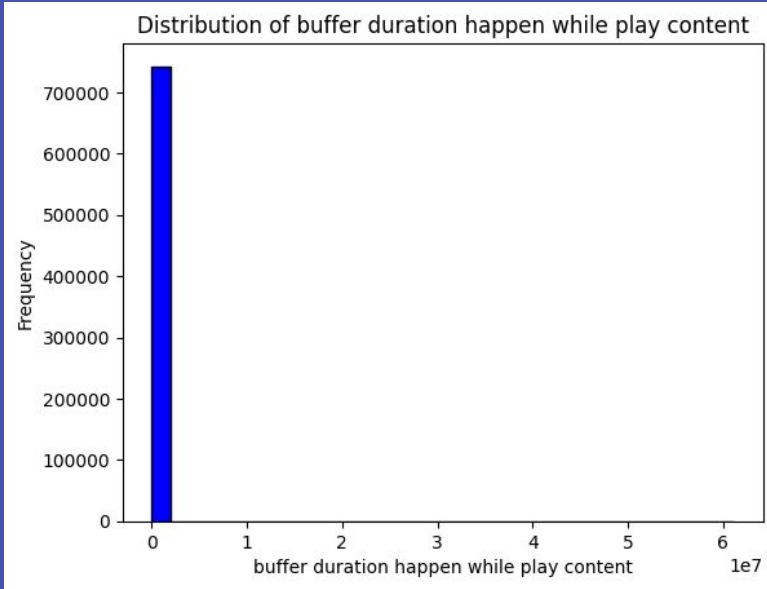
Distribution of Bytes When Play



The peak at the beginning indicates that most of the playback sessions may be very short or at very low quality (low bitrate). There may be technical issues such as frequent buffering or playback that stops before completion. This could be an indicator of a problem with the quality of service or the user's device.

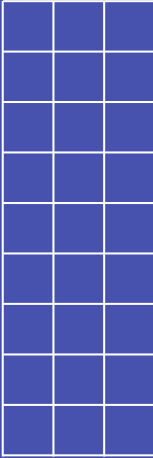
This distribution may also reflect user habits. For example, many users may only watch short clips of videos or prefer to play videos at low quality to save data.

Distribution of Buffer Duration

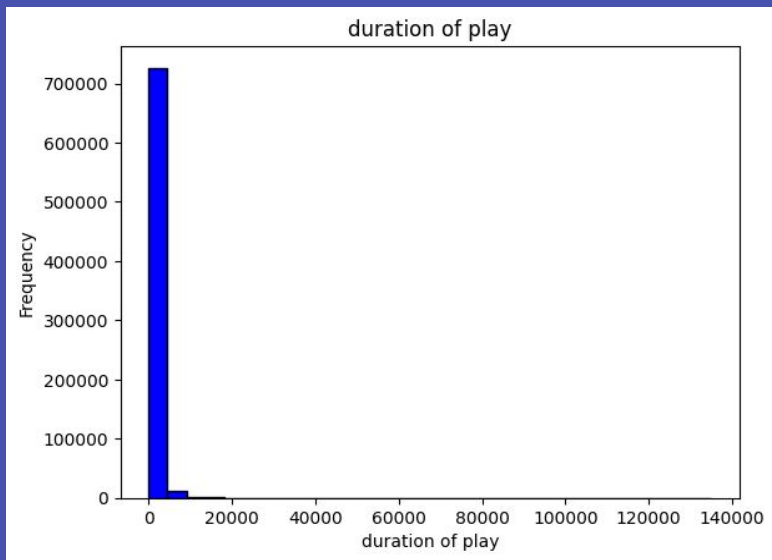


From the visualization results, it can be concluded that most users experience smooth playback without much buffering interruption.

By analyzing the distribution of buffering duration, we can identify technical problems that may occur on the server or network.

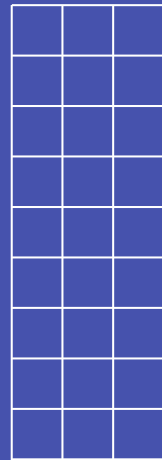


Distribution of Play

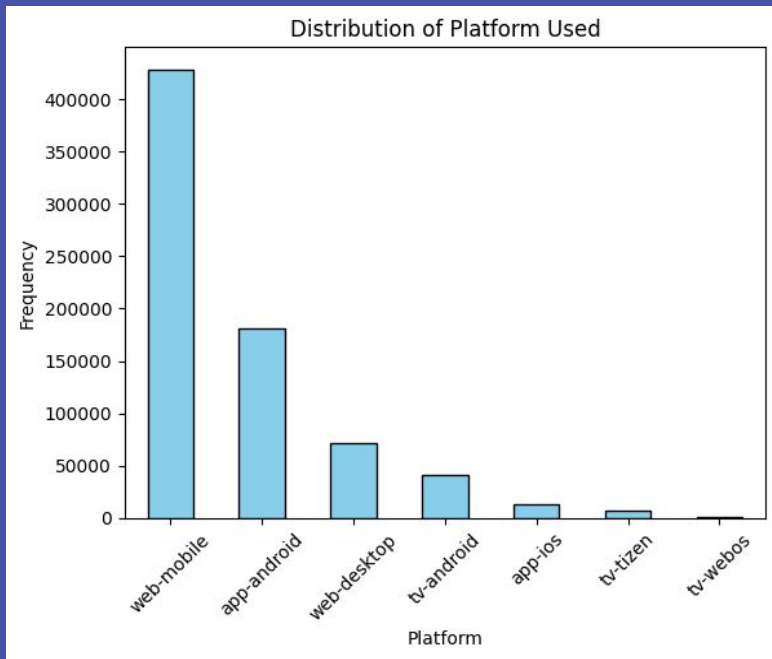


From the visualization results, it can be concluded that most users tend to play content in short durations. This could be because they are looking for quick information, or just want to see certain clips.

This distribution could also reflect user habits, many users may only watch short video clips or prefer to play videos with shorter durations.

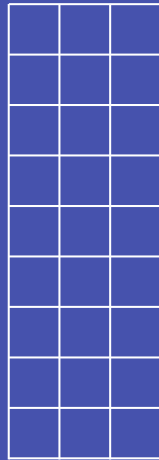


Distribution of Platform Used

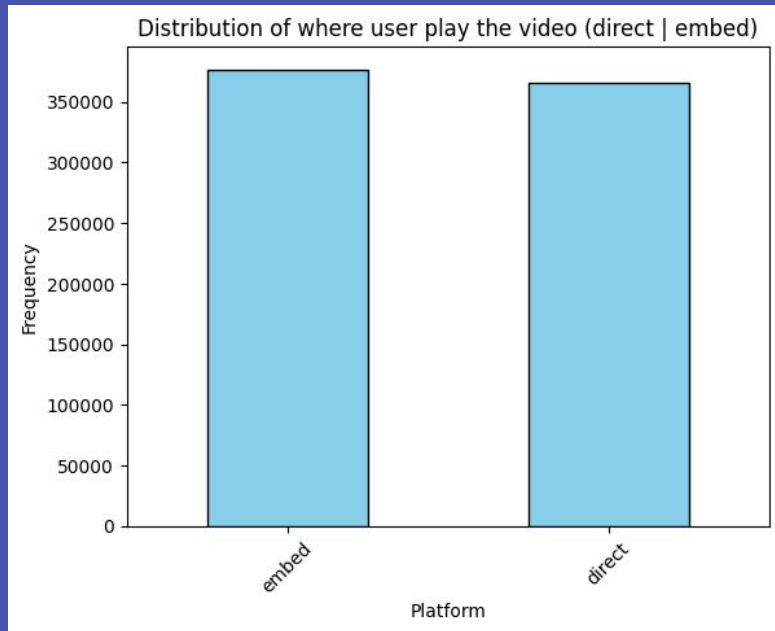


The "web-mobile" platform has a much higher frequency of use compared to other platforms. This shows that most users access services through mobile devices such as smartphones.

Users prefer to access services through mobile devices. This indicates the importance of optimizing user experience on mobile devices, especially on mobile browsers.



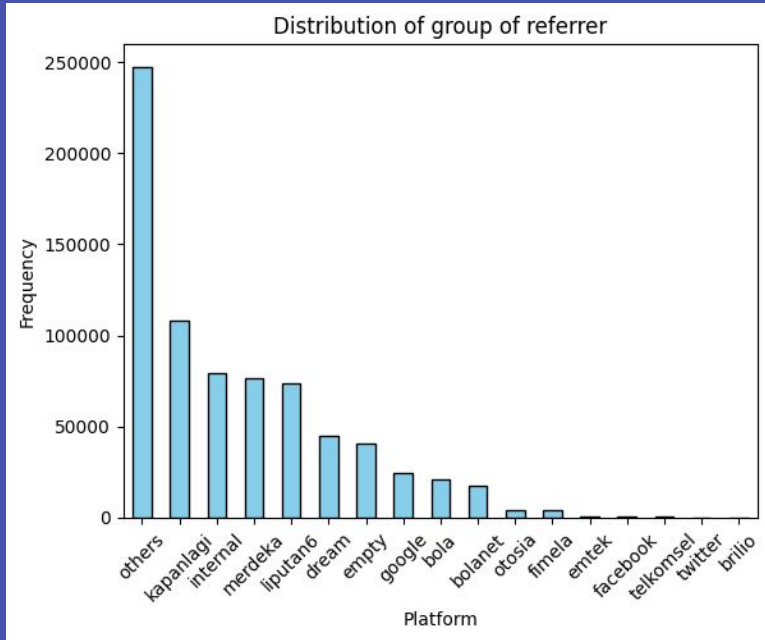
Distribution of Video Play (Direct vs. Embed)



The frequency of playback via embeds being almost the same as direct playback indicates that many users access the video through other platforms or websites that embed the video. This could indicate that the video embedding strategy has succeeded in attracting a large audience.

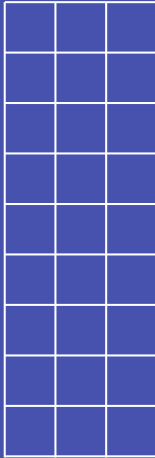
By analyzing further, we can find out the characteristics of users who prefer direct or embedded playback. This could help in market segmentation and content personalization.

Distribution of Group Referrer

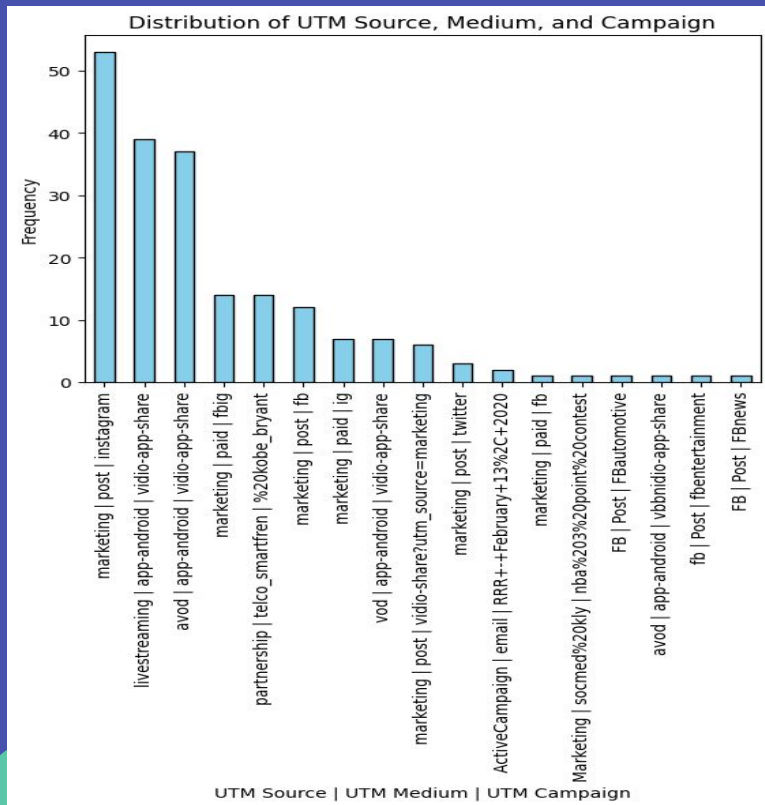


From the visualization, we can see that "others", "kapanlagi", and "internal" are the three largest referral sources. This means that most visitors come from these sources.

Since "kapanlagi" and "internal" are the main sources of referrers, the focus of cooperation can be prioritized on maintaining and improving relationships with these sources.

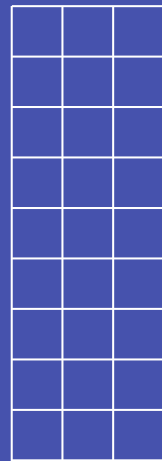


Distribution of UTM Source

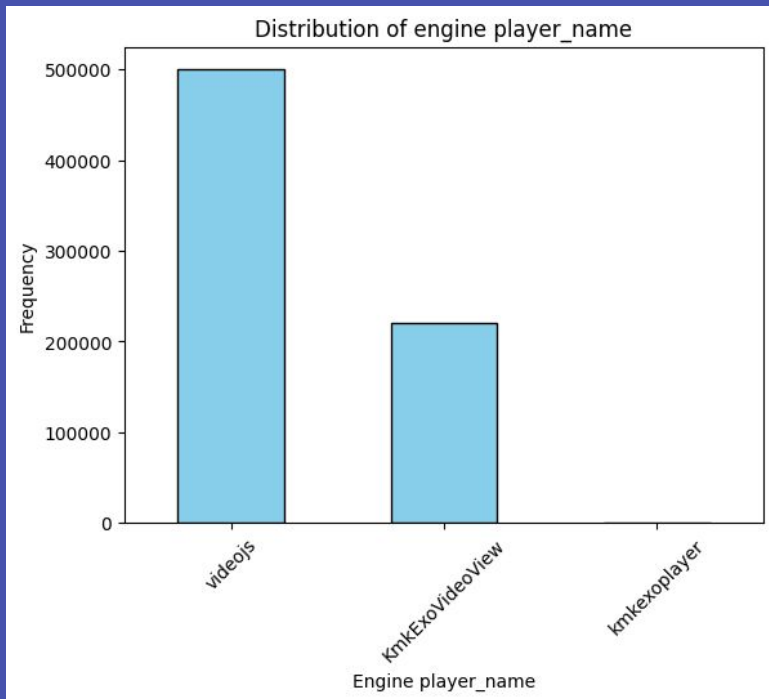


The combination "marketing | post | instagram" has the highest frequency. This shows that marketing campaigns on Instagram with post content type are the main source of traffic.

Some UTM combinations have low frequency. This could be an opportunity to optimize less effective campaigns or try new channels rather than maintain them.



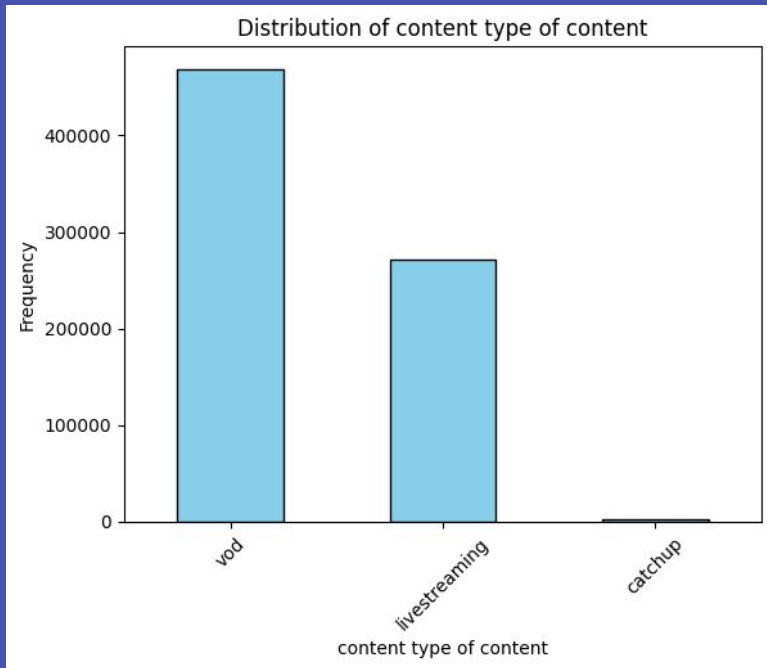
Distribution of Engine Player



Videojs is the most popular and widely used player engine. This could be because videojs has more complete features, is easier to use, or is more compatible with various platforms.

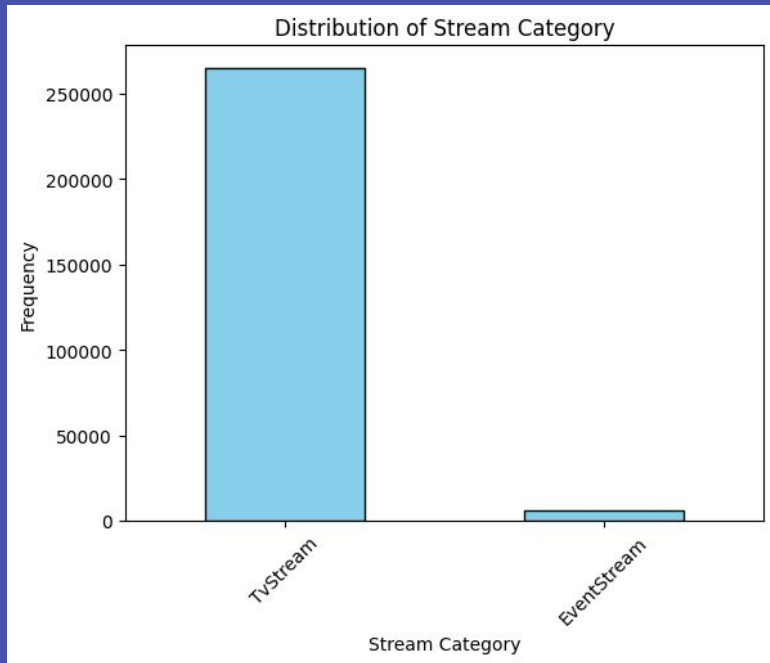
If we want to develop or optimize a video playback system, we should focus on videojs. This will allow us to take advantage of existing features and get support from a larger community.

Distribution of Content Type

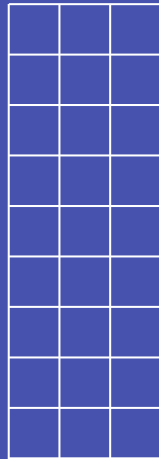


VOD is the highest in terms of content consumption. This shows that flexibility in choosing viewing time is very important for users. Meanwhile, live streaming has its own appeal, especially for real-time content such as sports, concerts, or news. However, the potential for live streaming is still great. Companies can increase live streaming events and improve broadcast quality to attract more viewers.

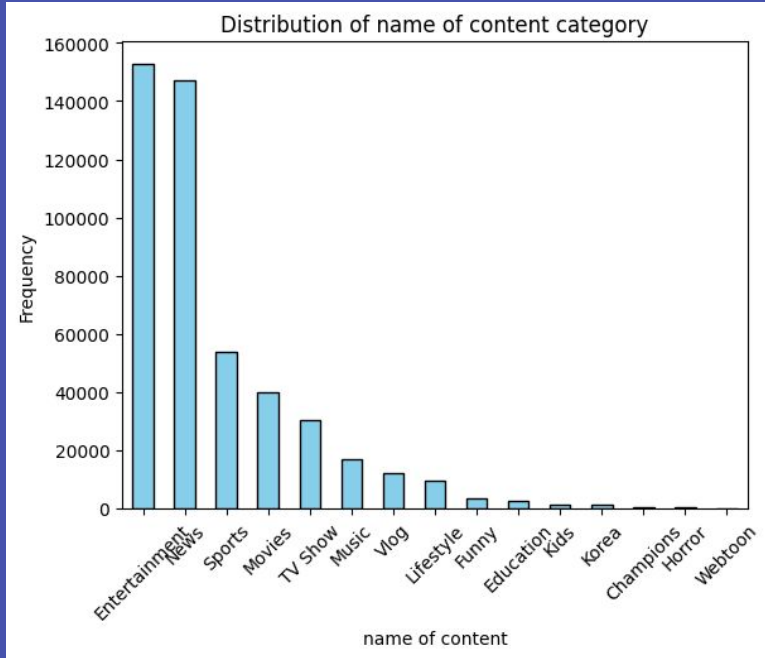
Distribution of Stream Category



TvStream is clearly the top choice for users. This could indicate that content such as TV shows, movies, or series are the most searched and consumed. However, EventStream has the potential to grow if improvements are made to the quality of content, promotion, and relevance of content to user interests.

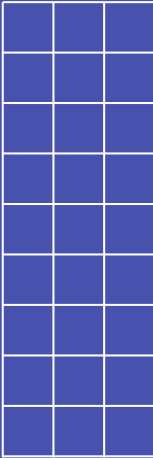


Distribution of Content Category



The "Entertainment" category has a much higher frequency of access compared to other categories. This indicates that most users are more interested in entertainment content such as movies, TV shows, or music.

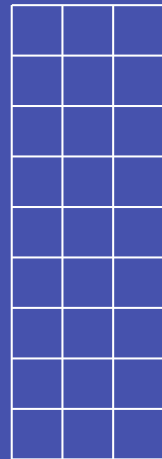
Although entertainment dominates, we need to consider developing content in other categories such as education, children, or Korean to reach a wider audience.



Insight & Story of Correlation Analysis



- The correlation between Average Bitrate and Bytes Used (0.113) shows a very weak relationship between the average bitrate and the number of bytes used during playback. This means that changes in the average bitrate do not significantly affect the amount of data used. Although bitrate affects video quality, modern video compression algorithms are quite efficient in managing bitrate without causing a drastic increase in file size.
- The correlation between Buffer Duration and Duration of Play (0.0063) shows no significant relationship between buffer duration and playback duration. This means that the frequency and duration of buffering are not affected by the length of the video played. Buffer duration is more affected by internet connection speed, network quality, and server capacity. High video quality or high bitrate can cause buffering to occur more often, but it does not always correlate with playback duration.



Statistics by Categorical Column #1

Based on the available data, it can be concluded that embedded playback offers a better viewing experience for most users, especially those with limited internet connections. However, if video quality is a top priority, direct playback may be a better choice.

Playback Location	Average Bitrate	Total Bytes	Buffer Duration
Direct	225985.805	59276020972240	820.357
Embed	176052.601	429450738876	6.103

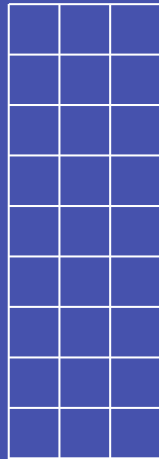
Statistics by Categorical Column #2

Platform	Play Duration	Total Bytes	Average Bitrate
App-Android	885.267	32126093775000	300000
App-Ios	529.009	209003777025	113914.308
Tv-Android	1177.327	17686420237500	300000
Tv-Tizen	1064.867	2517907200000	300000
Tv-Webos	0.963	42187500	300000
web-desktop	630.542	3031090640423	48855.194
web-mobile	125.383	4134913893668	174555.16

A Venn diagram with two overlapping circles. The left circle is teal and labeled 'The Best of the Best'. The right circle is pink and labeled 'The Best of the Worst'. The overlapping area in the center is shaded red.

TV platforms (Android, Tizen, WebOS) generally have longer playback durations and larger total bytes, indicating that TV users tend to watch longer videos. Mobile platforms (Android, iOS) have more varied durations and total bytes, possibly due to more diverse usage (e.g., watching short videos, live streaming).

Each platform offers a different viewing experience, both in terms of duration, quality, and type of content consumed. We need to optimize video quality and user experience for each platform. For example, on mobile platforms, focus on fast and efficient video playback, while on TV platforms, focus on high video quality.



Statistics by Categorical Column #3

Catchup content has the longest average playback duration, indicating that users tend to watch delayed content for longer periods. Each content type has different user characteristics. Catchup users tend to watch for longer periods and may prefer content that is already familiar. Livestream users tend to be more interactive and appreciate real-time content. VOD users are more flexible in choosing when to watch and may be more concerned with video quality.

Content Type	Play Duration	Total Bytes	Buffer Duration
catchup	1823.958	166880478456	49.182
Live streaming	745.144	55885305234957	1085.174
VOD	244.173	3653285997703	17.451



02

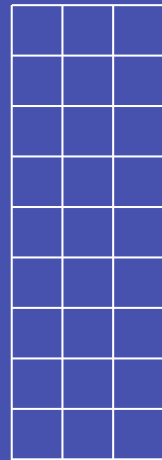
Machine Learning Model

Random Forest vs XGBOOST

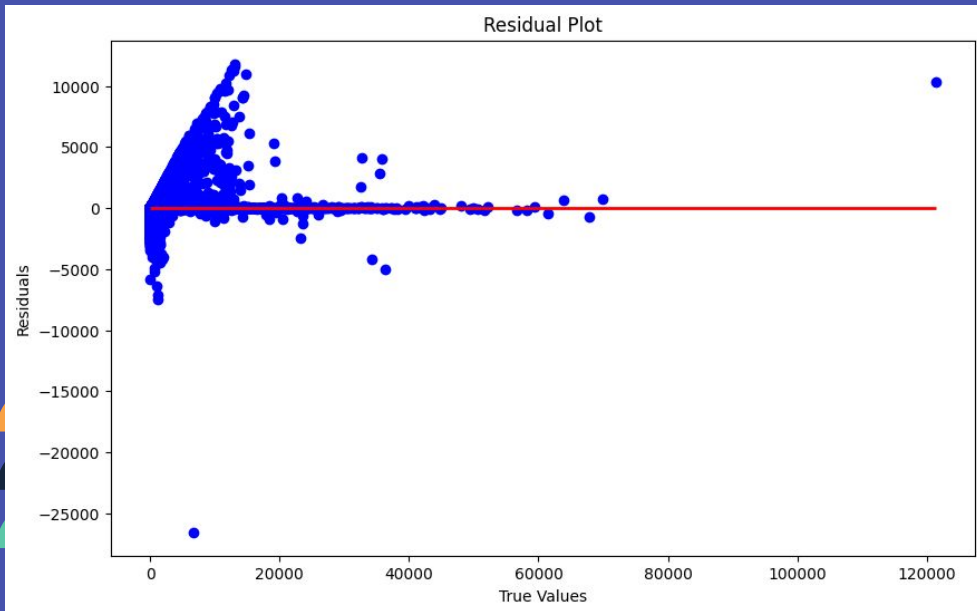
Summary

Random Forest: Based on the MSE value (664204.882), the Random Forest model has a much better performance than XGBoost in predicting playback duration. This means that the Random Forest model is able to produce more accurate predictions.

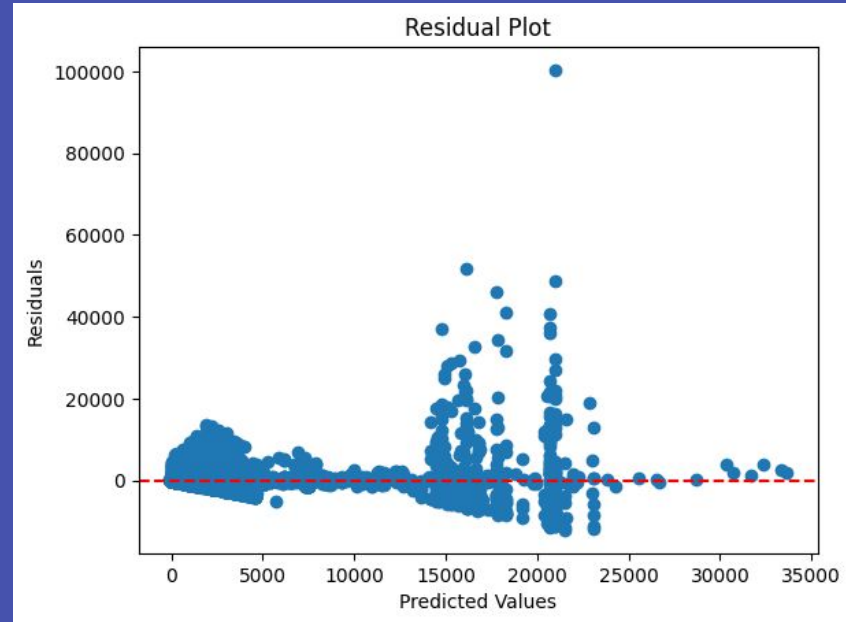
Potensi Overfitting: The very high MSE value (664204.882) in XGBoost indicates the possibility of an overfitting model. The XGBoost model may be too complex and overfit the training data, making it less able to generalize to new data.



Residual Plot #1





Random Forest




XGBoost

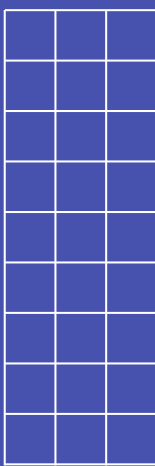
Residual Plot #1





Random Forest: In general, the residual distribution in Random Forest looks quite random around the zero line. This indicates that the model has been quite good at capturing patterns in the data. There are some outliers at the bottom of the graph, indicating that the model has difficulty predicting very low values. There is no clear heteroscedasticity pattern in the Random Forest graph.



XGBoost: The residual distribution in XGBoost is also quite random, but there are some patterns that are worth noting. There is a tendency for the XGBoost model to over-predict (predicted values are higher than actual values) for higher actual values. There are a few outliers, especially at the top of the graph, indicating that the model has difficulty predicting very high values. There is a slight indication of heteroscedasticity, especially on the right side of the graph, where the residual distribution widens as the predicted values increase.

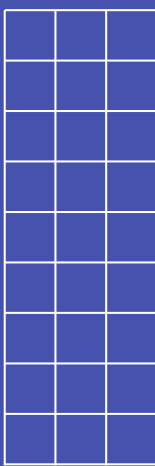



Residual Plot #2



Random Forest: Overall, Random Forest performs better in predicting data. This model is able to capture patterns in the data well and has few problems with heteroscedasticity.

XGBoost: XGBoost still has some shortcomings, such as a tendency to over-predict at high values and a slight heteroscedasticity. This indicates that the XGBoost model may need further tuning or consideration of using regularization techniques to reduce overfitting.





03

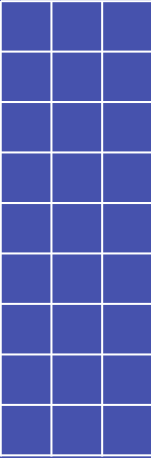
TOOLS

What I use and why?

Google Colab

0

I used Google Colab for all the activities because it offers a powerful cloud-based environment that allows seamless execution of Python code without the need for complex local setup. Its significant advantages include access to high-performance computing resources, such as GPUs, which are essential for machine learning tasks, and the ability to easily collaborate and share notebooks. Additionally, Google Colab provides an interactive interface that supports real-time visualization, making it ideal for both exploratory data analysis and building machine learning models.



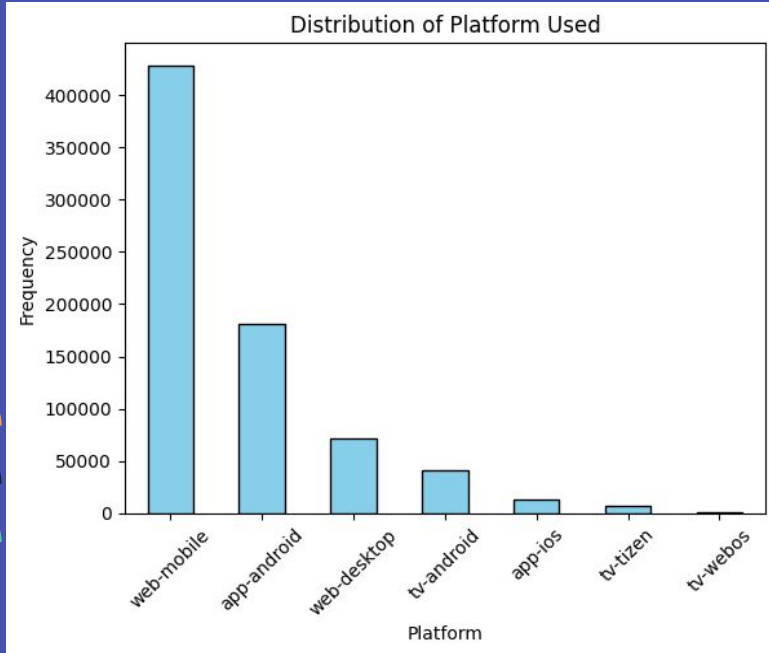


04

PLATFORM TO WATCH

By Analysis

Distribution of Platform Used



Based on the visualization results, it is clear that users show a strong preference for the "web-mobile" platform. However, a deeper understanding of user behavior will require consideration of additional factors such as platform availability, user demographics, and content types.

In my opinion, the reason users prefer to use web-mobile is because it is more practical without installing the application, it can be accessed directly through a browser, therefore we need to improve the quality of the web mobile platform to retain users so that they do not move to other platforms and do not forget to advertise the company's application.



05

TOP 10 VISITORS

Based on play duration per day.
[FUN QUERY]

TOP 10 VISITORS

By Play DURATION Per Day

No.	Watcher Id	Play Date	Play Duration
1	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	06/02/2020	638250
2	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	05/02/2020	621178
3	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	04/02/2020	530939
4	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	11/02/2020	514978
5	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	03/02/2020	497254
6	5d3dc12fb793d638ba679872d3b064c64e458455e971829c566998b7e3582e0d	16/02/2020	30840
7	4d3e14f878bc0d11d569732dc7117b2d04eafc304604e3fc31a980aedbca0366	16/02/2020	30180
8	0e8e3ed229cf18ad9a0fd8302699a4627fcd9e6c7a9aea6e2b6555490b456b3	01/02/2020	28050
9	27b6f6afe0bfaed18706af84b211ccb04d25ae73597ff5df686d437a3de6ad82	01/02/2020	27840
10	e2990e2b6c4532957cb25755115fcfdf4e1967776c641ad84736e0064eabf0bd	01/02/2020	27570

The background is a solid blue color. On the left, there is a large pink circle with an orange vertical bar passing through its center. The bar has a light blue section at the bottom. At the top center, there are three concentric white circles. On the right, there is an orange rounded rectangle with a green circle and a pink circle overlapping it. Below these, there is a 4x4 grid of small white dots. In the bottom right corner, there is a light blue stepped shape. At the bottom of the image, there is a green wavy line.

THANKS!

Thank you for the opportunity.