Unlocking YouTube Channel Performance Secrets

Aim:

This project aims to analyse YouTube channel performance by leveraging extensive metrics and using Machine Learning techniques to uncover patterns, trends, and actionable insights. We will focus on Exploratory Data Analysis (EDA), data visualization, and developing a predictive model to estimate revenue or subscribers based on the provided dataset.

Problem Statement:

Perform Exploratory Data Analysis (EDA), data visualization, and developing a predictive model to estimate revenue or subscribers based on the provided dataset.

Dataset Link

Step-by-Step Workflow

1. **EDA**

   i. Import required libraries

   ii. Load and explore the dataset

   iii. Handle missing values

   - Dropping few columns, as all their values are zero
   - Filled some columns with median
   - Dropped duplicate rows

   iv. Summary Statistics

   - Info of dataset

- Described float, int, object type columns

v. Handling Outliers

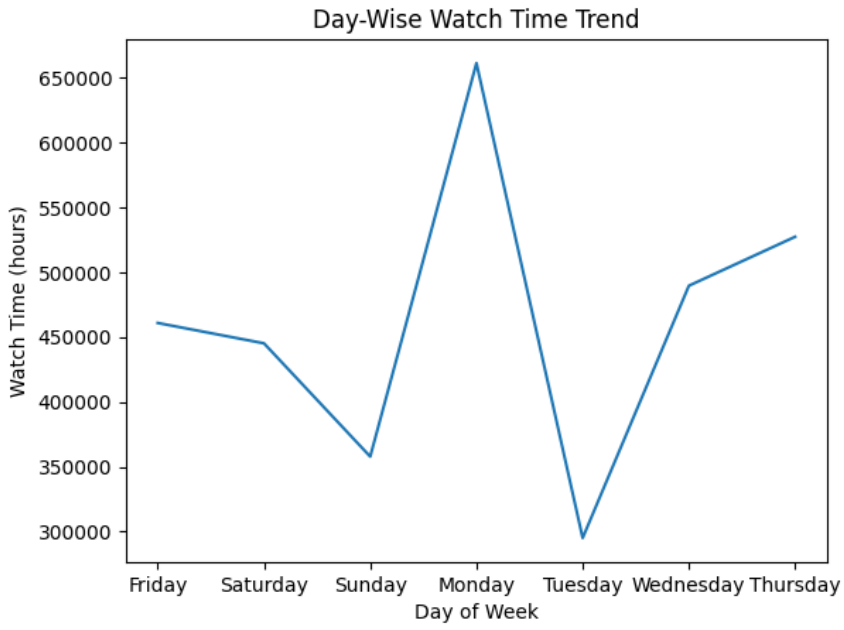
- Remove outliers using IQR technique

vi. Feature Engineering

- Encoded column ‘Day of Week’ (Friday=0)
- Convert 'Video Publish Time' to datetime
- Create revenue per view
- Create engagement rate

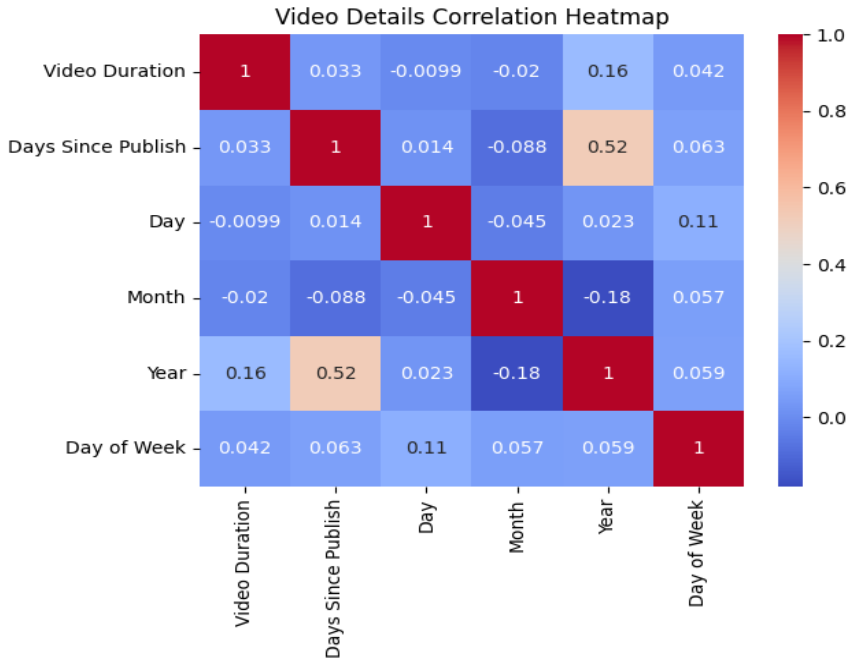
2. **Data Visualization**

i. Day-Wise Watch Time Trend



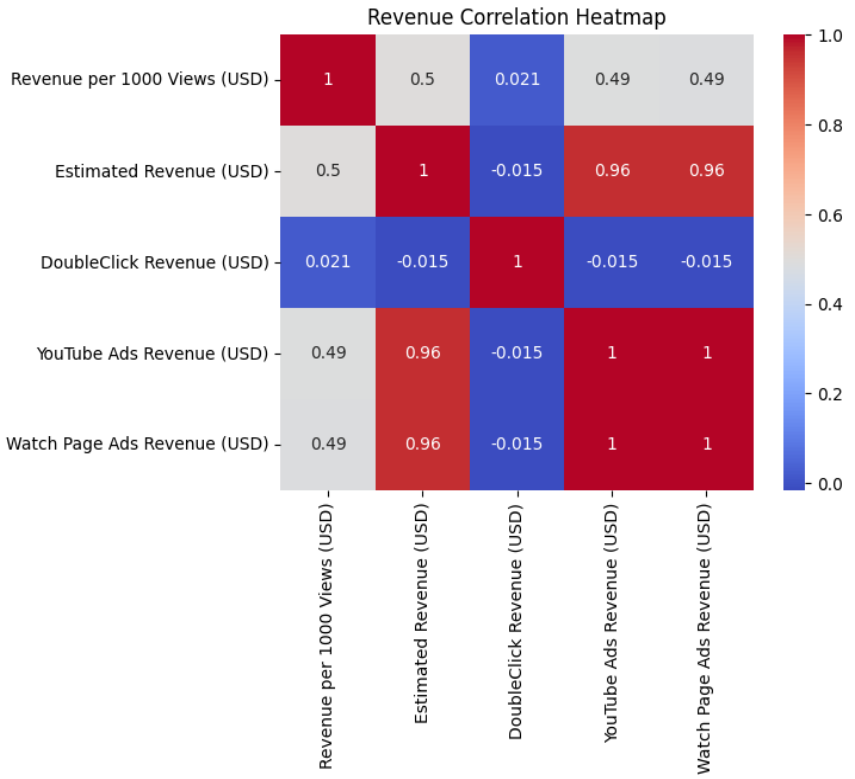
- Monday has max Watch Time (hours)

ii. Video Details Correlation Heatmap



- Weak Positive Correlation between "Video Duration" and "Year" (0.16)
- Insight: There is a slight tendency for videos published in more recent years to be longer in duration.

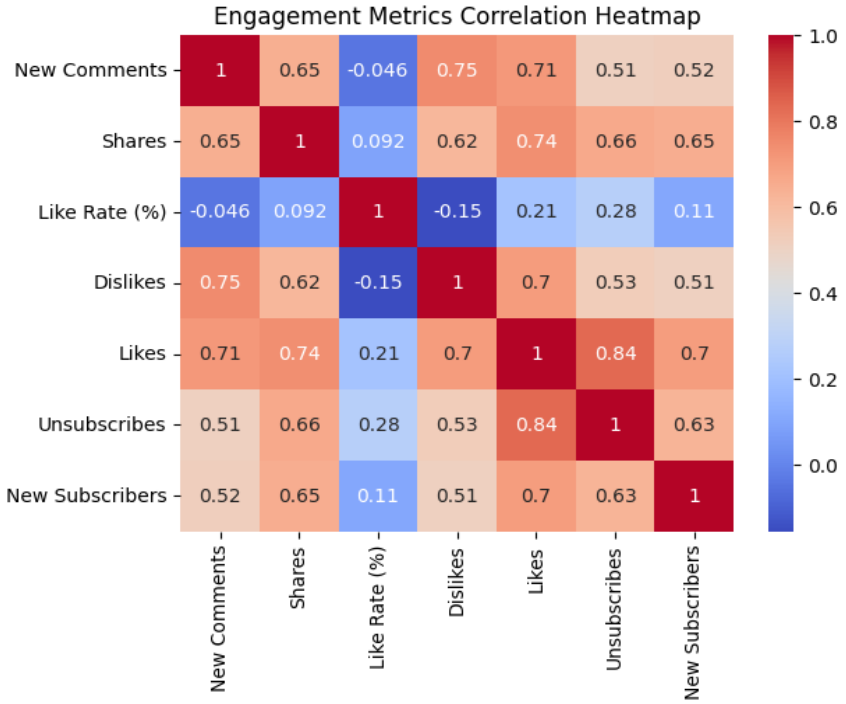
iii. Revenue Correlation Heatmap



- Estimated Revenue (USD), YouTube Ads Revenue (USD), Watch Page Ads Revenue (USD) have strong positive correlation
  - ✓ **Insight**: "Estimated Revenue" is driven by "YouTube Ads Revenue" and "Watch Page Ads Revenue."

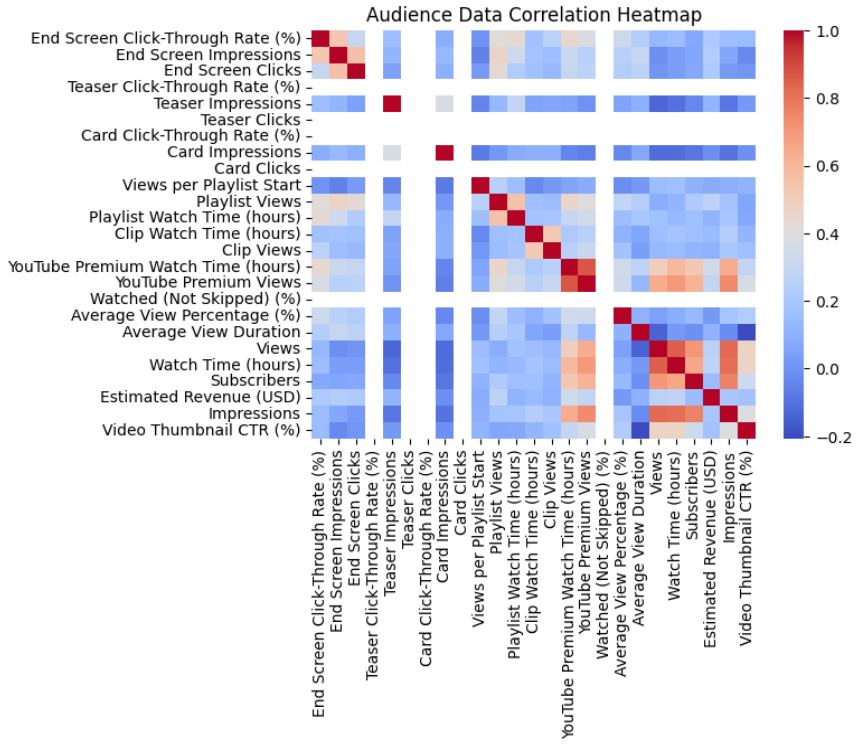
- Moderate Positive Correlation between "Revenue per 1000 Views (USD)" and "YouTube Ads Revenue (USD)", and "Watch Page Ads Revenue (USD)"
  - ✓ **Insight**: Revenue per 1000 Views is generated by ads
- Very Weak (Near Zero) Correlation with "DoubleClick Revenue (USD)"
  - ✓ **Insight**: "DoubleClick Revenue (USD)" shows almost no linear correlation with any of the other revenue metrics (values are very close to 0, like 0.021, -0.015).

iv. Engagement Metrics Correlation Heatmap



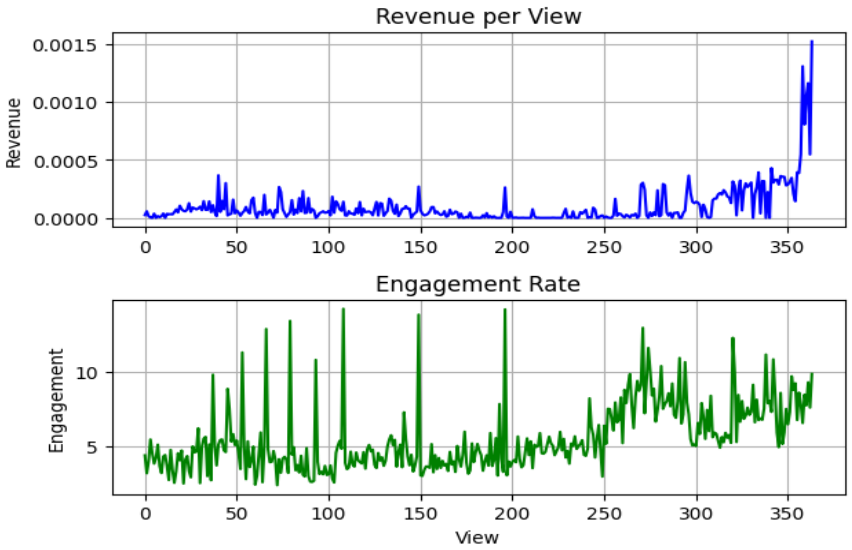
- Shares, New Comments and New Subscribers increases along with Likes
- New Subscribers is directly proportional to Dislikes/Share
- Likes and Unsubscribes (0.84) is a surprisingly strong positive correlation.
  - ✓ **Insight**: The video might be highly engaging but also controversial. Videos that perform well (get many likes) might be pushed to a wider audience, some of whom are not the target demographic and thus unsubscribe.

v. Audience Data Correlation Heatmap



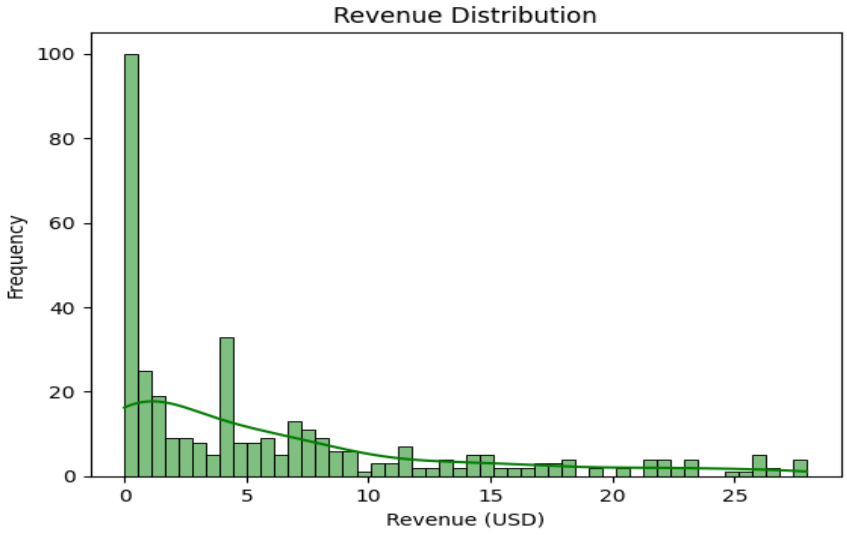
- End Screen Click-Through Rate (%), End Screen Impressions and End Screen Clicks are highly correlated among themselves. This is expected, as clicks are a function of impressions and CTR.
- Teaser Click-Through Rate (%), Teaser Impressions, and Teaser Clicks also show strong internal correlations.
- Card Click-Through Rate (%), Card Impressions, and Card Clicks are strongly correlated internally.
- Playlist Watch Time (hours), Clip Watch Time (hours), YouTube Premium Watch Time (hours), Watch Time (hours),
- Average View Duration", and "Views" are all highly positively correlated.
  - ✓ **Insight**: This indicates that videos with more views generally accumulate more watch time across various categories and have longer average view durations.
- Longer watch times are strongly associated with higher revenue and lead to more subscribers.
- A higher click-through rate on the video thumbnail means more people are clicking to watch the video after seeing it.
- This directly leads to more views, which in turn contributes to more subscribers and revenue. This highlights the
- importance of an engaging thumbnail for initial audience attraction.

vi. Revenue and Engagement Rate per View



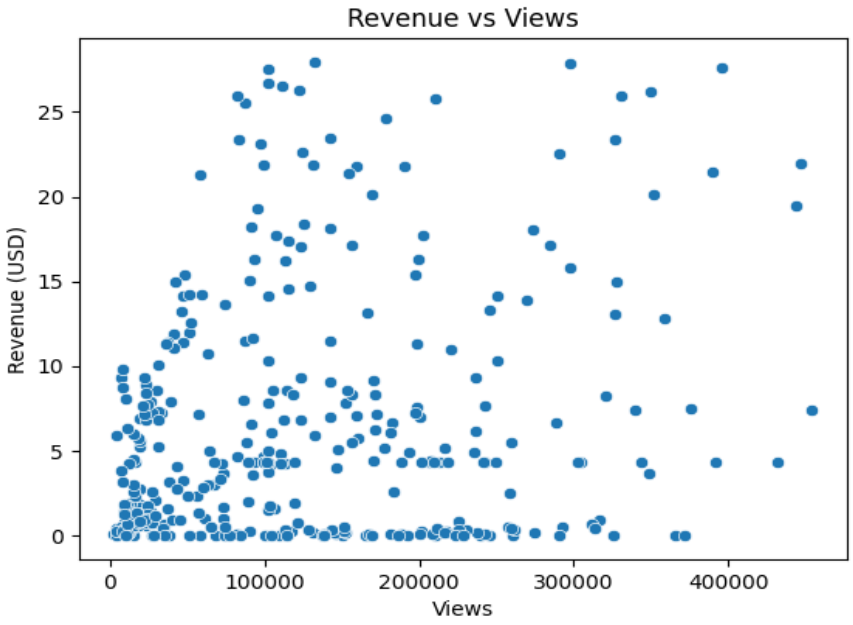
- The data indicates that as content accumulates more views, there is a tendency for both the engagement rate and the revenue generated per view to increase, particularly in later stages.

vii. Revenue Distribution



- The revenue distribution is highly unequal, with a large volume of content generating very little income, and a small fraction of content responsible for most of the higher earnings.

viii. Revenue vs Views



- While more views generally lead to more revenue, the relationship is not perfectly linear. Other factors significantly influence how efficiently views are converted into revenue, leading to a wide range of earnings even for content with similar view counts. There is a clear segment of content that achieves high views but yields minimal revenue.

ix. Top Performers by Revenue

ID	Estimated Revenue (USD)	Views	Subscribers	
297	297.0	27.955	131817.0	232
157	157.0	27.882	298148.0	275
240	240.0	27.575	396276.0	180
149	149.0	27.505	101950.5	180
196	196.0	26.727	101950.5	180

3. **Predictive Model**

i. Prediction using Linear Regression

- ✓ Import required libraries
- ✓ Define features and target variable
- ✓ Split the data into training and testing sets
- ✓ Initialize and train various models
- ✓ Make predictions
- ✓ Calculate the prediction accuracy and root mean square error

ii. RSME Results:

Model	**RMSE**
Linear Regression	0.0004752978199724935
Random Forest Regressor	0.5731469471739311
Decision Tree Regressor	2.0574488092634833
K-Neighbors Regressor	8.90953470728264
MLP Regressor	404.36641664541054

iii. Accuracy Results:

Model	**R^2 Score (Accuracy)**
LinearRegression	0.9999999974053332
RandomForestRegressor	0.9962270425146201
DecisionTreeRegressor	0.9513807761651111
KNeighborsRegressor	0.08828312673320471
MLPRegressor	-1877.0195262839095

Conclusion:

This project successfully explored the dynamics of YouTube channel performance using a combination of Exploratory Data Analysis (EDA), data visualization, and machine learning models. Through detailed analysis of 53 features, we extracted critical insights related to channel metrics and developed predictive models to estimate subscriber count or revenue.

To evaluate prediction accuracy, five regression models were trained and tested. The results are shown above.

The analysis successfully applied multiple machine learning models to predict YouTube channel performance. Among them, **Linear Regression** and **Random Forest Regressor** performed best, showing the lowest RMSE and highest R^2 scores, indicating strong predictive accuracy. In contrast, models like **MLP Regressor** and **KNN** showed poor performance, highlighting the importance of model selection based on dataset characteristics. This study confirms that simpler models can sometimes outperform complex ones when data is well-structured and clean.

These findings highlight the importance of selecting the right model architecture based on data characteristics. Linear models and tree-based ensembles performed exceptionally well in this case, offering accurate and interpretable results. This analysis provides a strong foundation for channel creators and marketers to make data-driven decisions and optimize performance strategies on YouTube.

Github Link

Colab Notebook Link

Power BI Dashboard Link