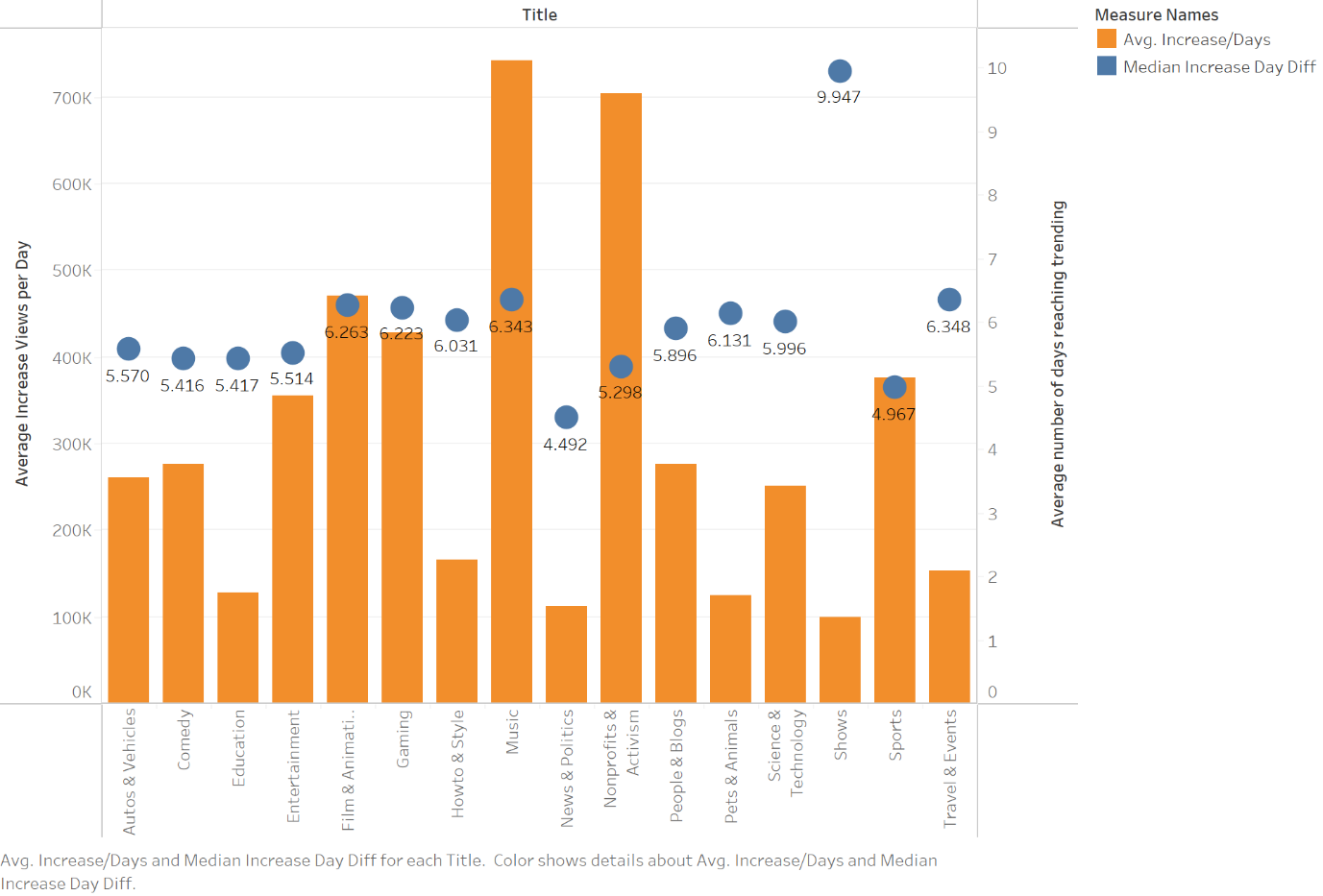
YouTube advertisement putting project report

1. Exploratory Data Analysis (EDA)
2. Put advertisements on what kind of videos?

Plot:

表格

描述已自动生成



Result:

Our app is related to travel, and travel & events has a relatively shorter time it takes to reach trending and the total views are not too high. The cost is acceptable.

Based on the form and plot, we can conclude that the shorter the time it takes to reach trending and the video category with the highest average daily views growth is more conducive to advertising. However, considering that videos with higher total views may have higher placement costs, I have made the following plan:

1. Music + Nonprofits & Activism:

Pros: The average daily viewing volume of these two types of videos has increased significantly, and the medium number of days a video reaches trending is relatively low.

Cons: The total viewing volume of music videos is the highest, so if advertising is placed, the cost may be higher. However, the total viewing volume of nonprofit videos is relatively low, and even if it grows rapidly, the audience may not be as broad as other categories, which may have an impact on the advertising effect.

1. Gaming + Comedy/ Entertainment + People & Blogs

Pros: The total viewing volume of these types of videos is not low, but because they are not the highest, they can be controlled in terms of cost. At the same time, their average daily traffic growth rate is also relatively high, and their spending time on trending is not particularly slow, which can be a good solution.

Cons: Considering that the main product is a travel app, choosing videos such as gaming and entertainment may not necessarily have a strong interest in tourism among their audience.

1. When is the best time to put advertisements?

By analyzing the time difference between views and video releases, as well as the trending time on the list, I believe that the appropriate time for advertising should be when views rise the fastest.

* By daily:

图表, 折线图

描述已自动生成

We can see that on the 6th of each month, the average daily growth views increase fastest which is around 700,000 views per day. Overall, the values of the 4th, 6th, 24th, and 29th are relatively high, making them suitable for advertising and promotion.

On the 28th, the average daily growth views increase lowest which is around 200,000 views per day. The values on the 15th, 22nd, and 28th are relatively low, so it is recommended not to advertise.

* By monthly:

We can see that in June, the average monthly growth views increase fastest which is around 770,000 views per day. Overall, the values from April to June are relatively high, making them more suitable for advertising. The best time is in June.

From July to October, the average monthly growth views are all relatively low, which should not be a good time to put an advertisement.

图表, 折线图

描述已自动生成

1. Data preprocessing

Correlation map:

图表

描述已自动生成

* Views和likes, dislikes, comments都呈正相关

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

* 三张散点图可以看出views与他们呈linear relationship

1. Check and remove fraud data

By using decision tree model to analyze the linear relationship between views and likes, dislikes, and comments, a threshold of 4 is set to filter out rows with abnormal views.

**Process:**

1. Divide the data into corresponding groups according to the video category and extract the training and testing sets separately.
2. Train each group separately using linear regression and decision tree models.
3. Predict the test set and calculate the error value using mse.
4. Comparing the size of mse between two models for each group, it was found that 93.75% of the groups performed better in the decision tree model.
5. Filter all data using a trained decision tree model – 145 rows
6. Remove data has views but no likes/dislikes/comments – 108 rows
7. Predict future views trends
8. Linear Regression

Train a linear regression model by comparing the time difference between trending date and publishing date with views, and then use 1, 15, and 30 as predicted days to add them to the model to draw a discounted trend graph.

* Others trend decreasing

Trending:

地图

描述已自动生成

Publish:

图表

描述已自动生成

publish

1. ARIMA

Train ARIMA models with trending date and publishing date and views respectively

* Publish Date

图形用户界面, 应用程序

描述已自动生成

* Trending Date

地图

描述已自动生成

1. Classification
2. Word2Vec

考虑到title, channel title, tags和description都是文字数据，需要进行文本数据处理。首先使用Word2Vec模型进行处理。

* 去除所有非英文字母的特殊字符，只保留英文字母
* 去除英文stopwords
* 所有文字转为小写

训练好后将title, channel title, tags和description合成一列作为classification model的input

* Logistic Regression

一些文字和图片的手机截图

描述已自动生成

* SVM

一些文字和图片的手机截图

中度可信度描述已自动生成

* Random Forest

一些文字和图片的手机截图

描述已自动生成

1. Clustering - Kmeans