Krikey EDA Report

**Date:** 25th April, 2022

**Place:** Toronto, Canada

**Introduction:**

This document provides a summary of **Part-1: Exploratory Data Analysis** carried out as part of Krikey’s take home assignment challenge. This report starts by providing a brief overview about the challenge, followed by details about the data sources that were provided, and their relationships identified among the data sources using Entity Relationship Diagram (ERD) diagram. Later, methodology followed during the data analysis is described. This is followed by outlining the data analysis findings using charts and/or tables. Integrating data analysis results with business outcomes is a crucial part and this is handled in the product recommendations section. Finally, remarks on the future data analysis that is proposed to address a set of outstanding questions are listed to initiate further discussions with the business team.

**Overview:**

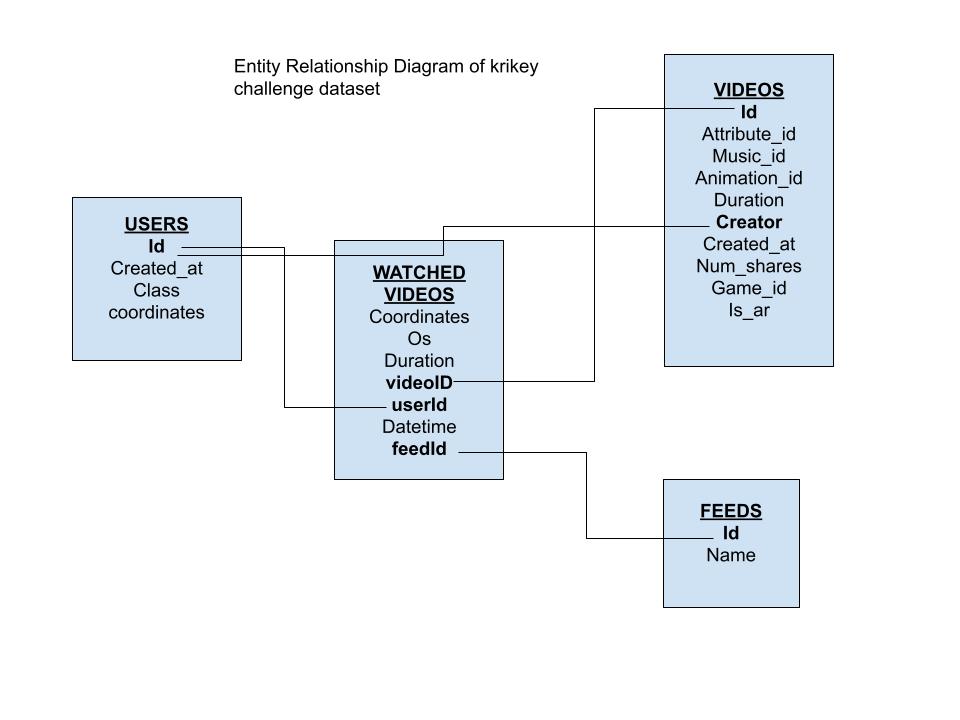
The requirement of this challenge is to perform exploratory data analysis using the data provided by krikey team, report interesting findings using either visuals or tables, and more importantly to outline the story behind the data that would be useful for the business development.

**Data Sources:**

Data provided for this analysis can be downloaded from [here](https://drive.google.com/drive/folders/1hSHQVl380yYS0CzIDzpYFnCvXHenXH2G). Data is provided in 4 different csv files. Each csv file has specific information as below:

1. **watchedVideo**: This contains all occurrences of the watchedVideo event.
2. **users**: This table contains the registered users
3. **videos**: This table contains all created videos
4. **feeds**: This contains the feeds the videos are served on

Preliminary analysis of the data suggested that the above datasets are connected using few unique identifiers. A quick pictographic representation of the connection between the datasets is shown below using a ERD diagram.



**Methodology:**

The methodology of the data analysis follows:

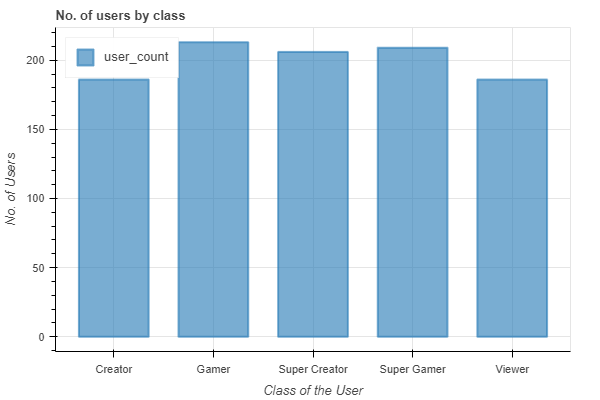
1. Asking a rigorous question
2. Analyzing or transforming the data concerning the question raised in **a.**
3. Interpreting the transformed data (**b.**) by creating interactive visuals or consolidated data tables
4. Documenting the findings and limitations of the analysis, if any.

**Results:**

This section provides results from the data analysis. As per our observations of the data, we organized the results into 3 subsections to provide a comprehensive picture about the datasets. First subsection outlines findings about “User Registrations” by geographical class, geographic location, and time. This is followed by remarks on “Video Creation” by datetime, location and other features. Finally, we present the results about “Videos watched” and their correlations with users. For every finding, the corresponding data analysis source code link is provided.

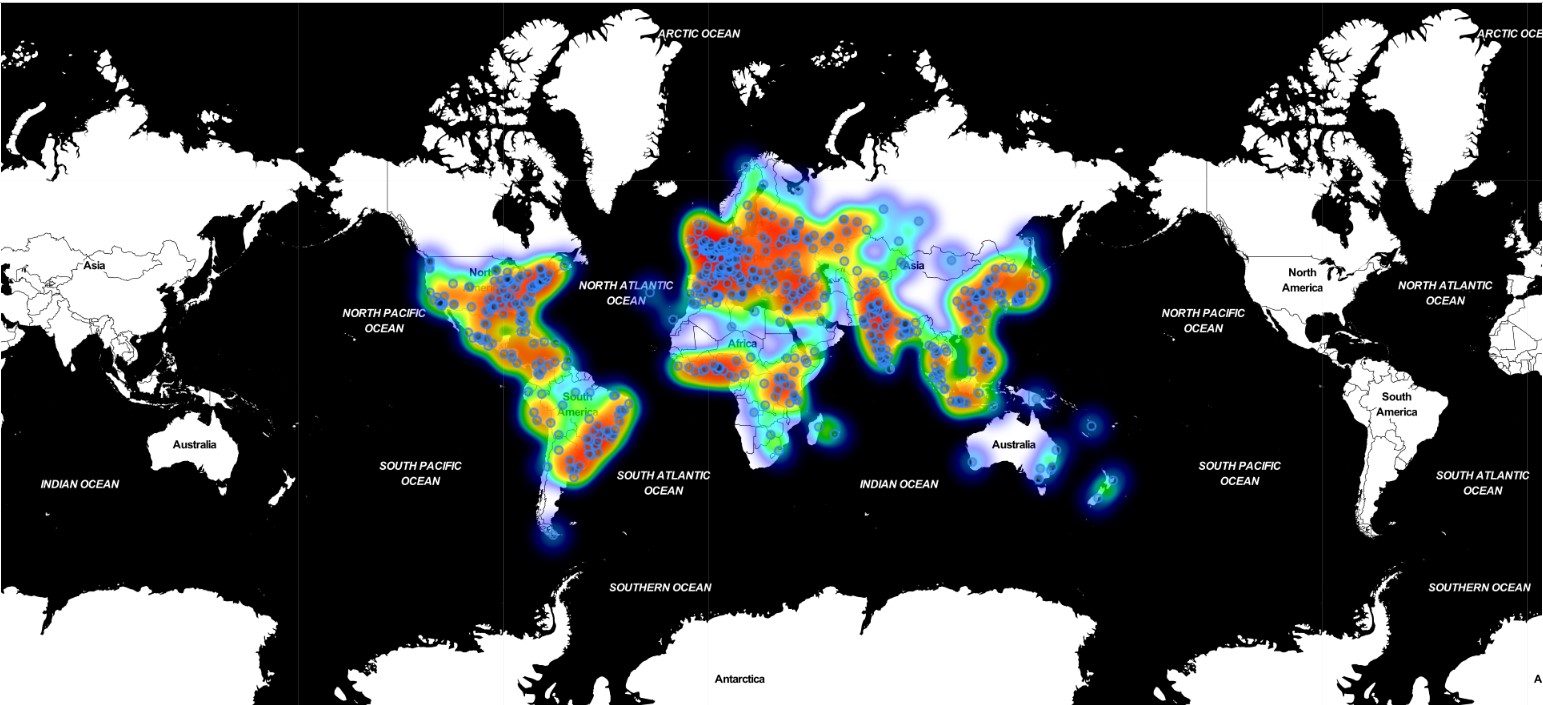
1. **User Registrations:**

Every user, of the total of 1000 users, belongs to one of the 5 classes. Below bar chart shows the number of users belonging to a specific class. Gamers class, with 213, is the top user class among 5 and Creators and Viewers classes, shared equally by 186 users, are among the 2 least popular classes.

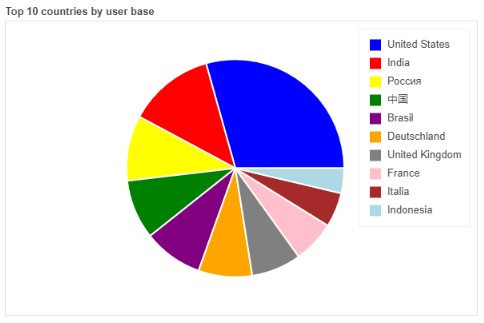
****

**Source:** Section 7.1.1 in 1\_exploratory\_data\_analysis\_krikey.ipynb

It is observed that users registered from places all around the world. This information about user registrations is shown using heatmap and points on the world map attached below. To identifying the top10 countries with large user registrations, we convert the latitude and longitude of registered user into country and then show the results in a pie chart attached.

****

**Source:** Section 7.1.3 in 1\_exploratory\_data\_analysis\_krikey.ipynb

****

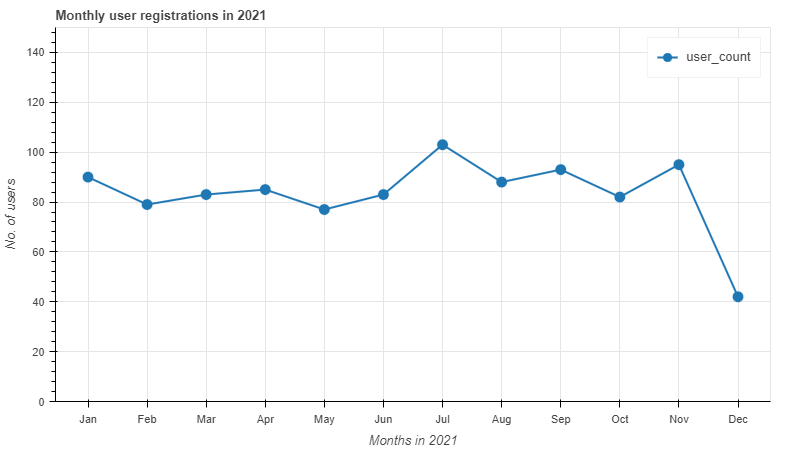
**Source:** Section 7.1.2 in 1\_exploratory\_data\_analysis\_krikey.ipynb

Further, an interactive multi-marker map has been implemented to display the concentration of registered users on the world map. A snapshot of the the interactive map is attached here.

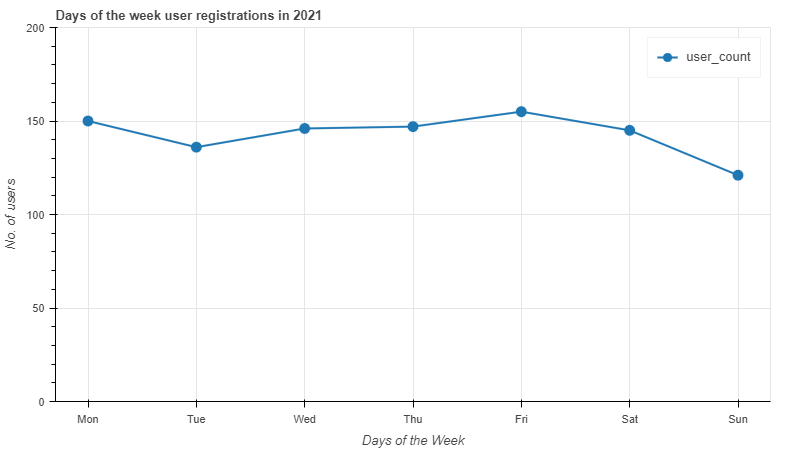


**Source:** Section 7.1.4 in 1\_exploratory\_data\_analysis\_krikey.ipynb

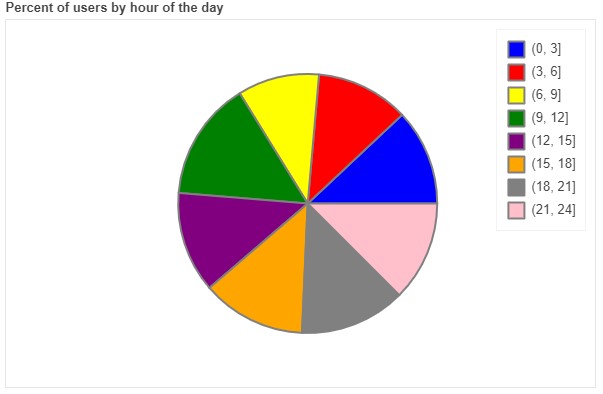
Change in user registrations over time is presented using two line charts. First line chart shows no. of user registration on a monthly basis in 2021. Except Dec 2021, user registrations remain close to an average level of 80 every month in 2021. To see the impact of day of week on user registrations, we plot no. of user registration for every day of the week in 2021. Interestingly, user registrations fell below the average value on Sundays. Another time aspect, hour of the day, has little to no impact on the no. of user registrations. This is depicted by the pie chart that is attached below.

****

**Source:** Section 7.1.5 in 1\_exploratory\_data\_analysis\_krikey.ipynb

****

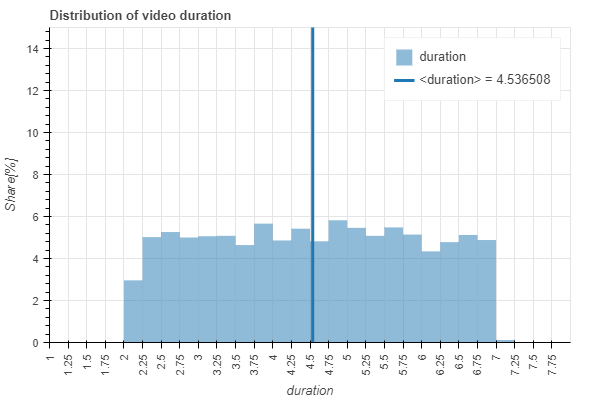
**Source:** Section 7.1.6 in 1\_exploratory\_data\_analysis\_krikey.ipynb

****

**Source:** Section 7.1.9 in 1\_exploratory\_data\_analysis\_krikey.ipynb

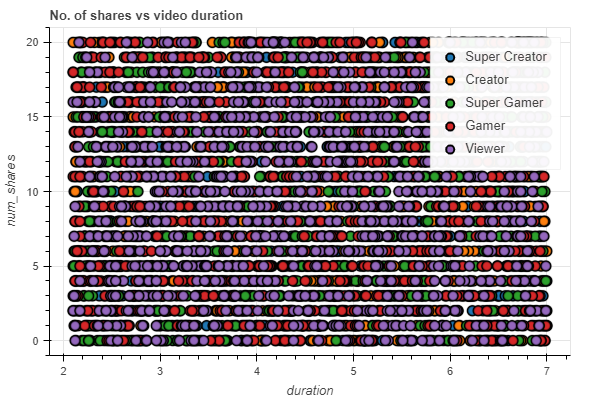
1. **Videos Creation:**

Video dataset has 5000 entries. This means 100 registered users created 5000 videos during 2021. Every video has a different duration. We look at the distribution of video duration to see if there is a pattern. Duration distribution looks uniform with min value of 2.1min, max value of 7min and an average duration of 4.5min.



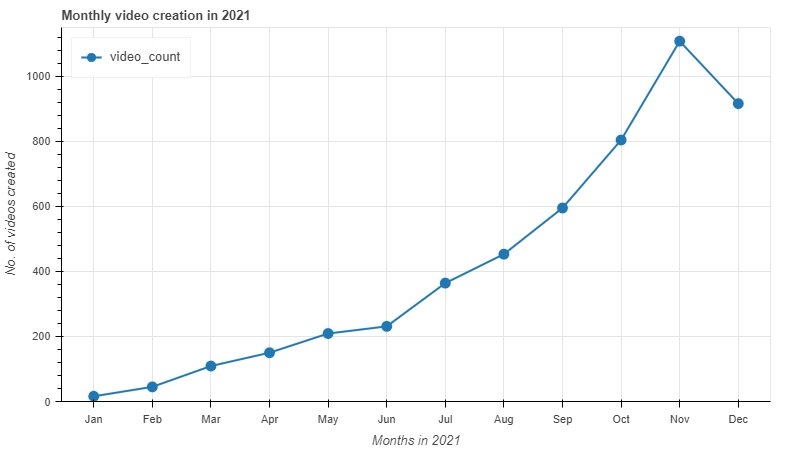
**Source:** Section 7.2.1 in 1\_exploratory\_data\_analysis\_krikey.ipynb

The average duration and average no. of shares of non-AR and AR video is pretty much **similar**. Further, the mean duration of the video created by users of different classes also remained close to ~4.5min. [**Source:** Section 7.2.2 and 7.2.3 in 1\_exploratory\_data\_analysis\_krikey.ipynb]. To identify if there is any correlation between no. of shares vs video duration, we create a scatter plot as shown below. There is **no** particular relationship that was found between num\_shares vs duration of videos.

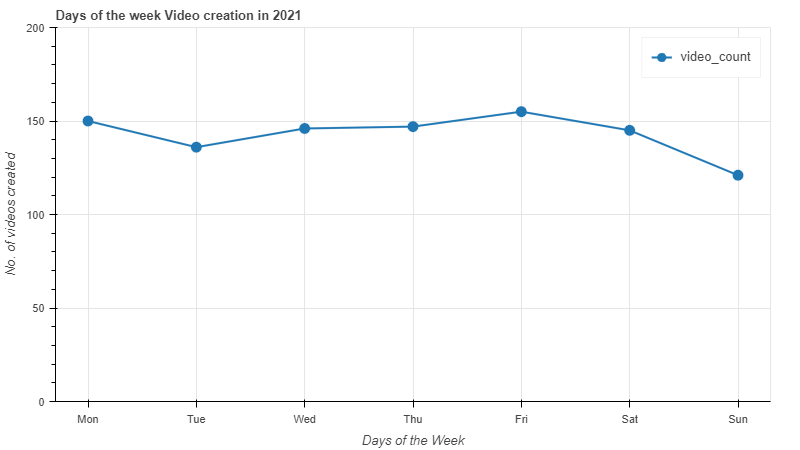


**Source:** Section 7.2.4 in 1\_exploratory\_data\_analysis\_krikey.ipynb

Further, to see how videos are created over time, we created a number of videos created on a monthly basis and also looked at no. of videos created during the day of week.



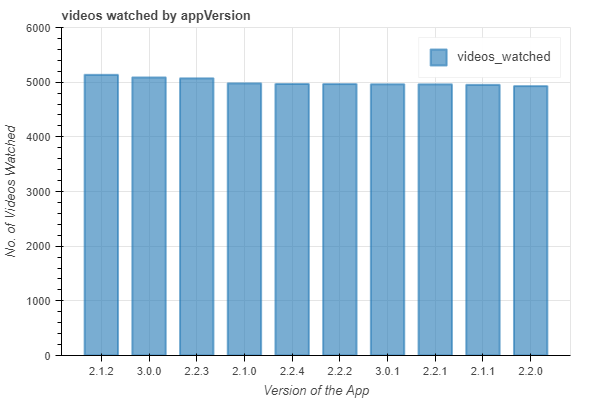
**Source:** Section 7.2.5 in 1\_exploratory\_data\_analysis\_krikey.ipynb

****

**Source:** Section 7.2.6 in 1\_exploratory\_data\_analysis\_krikey.ipynb

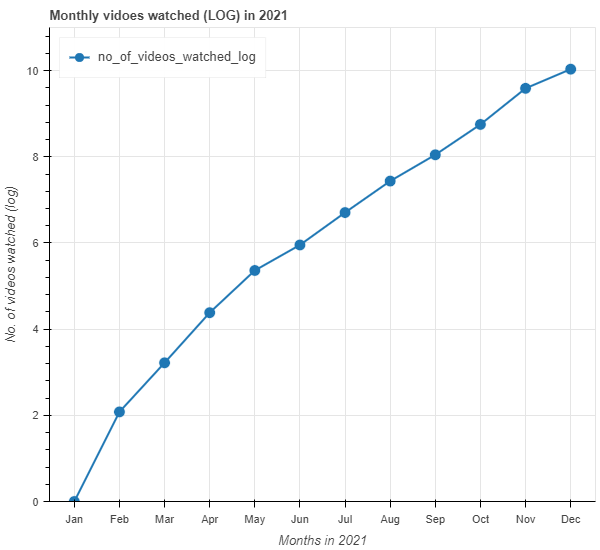
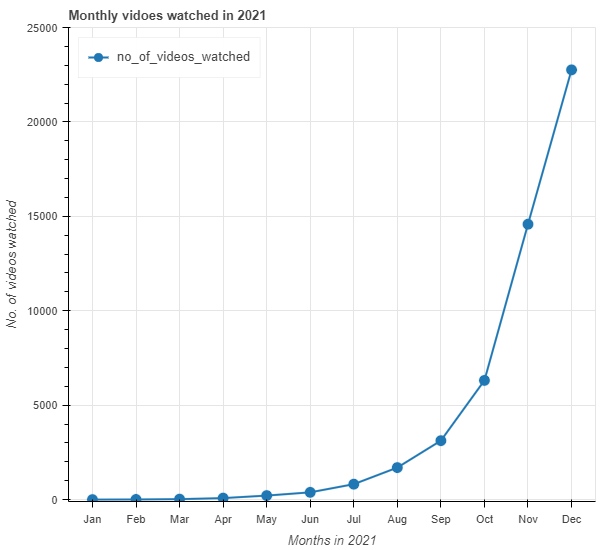
1. **Videos Watched:**

In this final section, we look at different attributes that contributed to the change in no. of videos watched. Starting our analysis by OS. Android and iOS neither of the platforms had a strong impact on videos watched. Whereas, the app version showed a small impact on no. of videos watched. Version 2.1.1 and 3.0.0 are the top 2 version with regards to no.of videos watched and the least performing one is version 2.2.0. This being said, we should remember that the average no. of videos watched on all version is very close to ~5000.



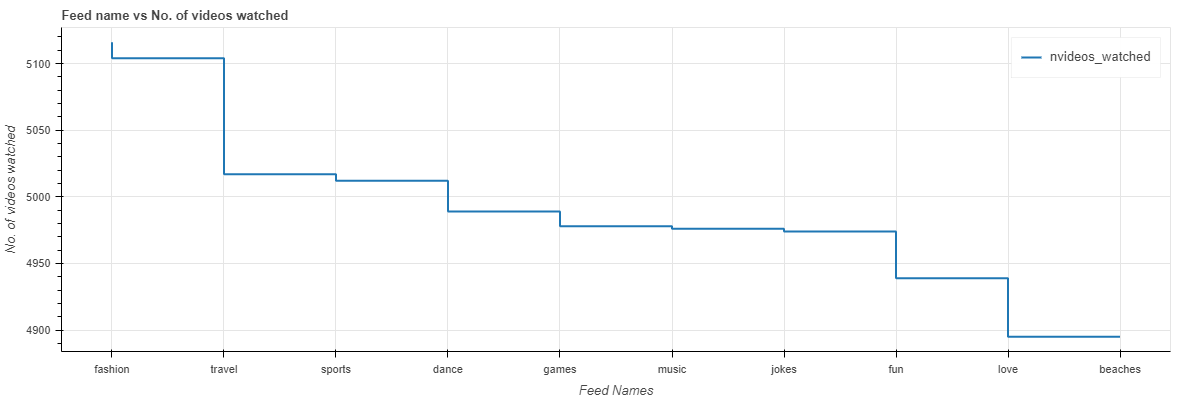
**Source:** Section 7.3.2 in 1\_exploratory\_data\_analysis\_krikey.ipynb

Month over month, no. of video watched increased exponentially in 2021. Below line plots show actual no. of videos watched on left and no. of videos watched in log scale on the right.



**Source:** Section 7.3.3 in 1\_exploratory\_data\_analysis\_krikey.ipynb

Below graph clearly shows the popular feed. Fashion is at the top, followed by travel, sports and beaches being the least popular feed.



**Source:** Section 7.3.4 in 1\_exploratory\_data\_analysis\_krikey.ipynb

**Product Recommendations:**

With our analysis, we arrive at below rudimentary product recommendation.

* The user registrations are higher in the USA, and followed by India. If we plot change in user registration over time, we can identify if there is an increasing or decreasing trend. This can lead us toward promoting our product in the countries showing a strong positive trend.
* User registration was observed to decrease in Dec2021 and especially on sundays. Identifying the reason behind the decrease will help us guide towards a stable growth.
* No. of videos watched increased exponentially until Nov2021 and dipped in Dec 2021. Identifying the type of videos watched in Dec 2021 compared to other months will give us an idea about the dip and also user behavior.
* There is clearly a strong inclination towards a particular type of videos based on the feed. Recommending user to create and share more videos in the same genre will fuel additional exponential growth.

**Limitations:**

The data provided is close to uniform, which indicates that the data might be simulated. Looking at a scatter plot between duration and shares or distribution of duration will confirm the limitations. With a uniform data set as provided, there is very less possibility to identify strong trends.

**Next Steps:**

Next stage of the analysis should address the remainder of outstanding questions. Few of which are listed below.

1. Location of user registration over time (monthly)
2. Location of users who contributed to the no. of views (top10)
3. Top 10 users who created more videos
4. Top 10 users whose videos are most shared.
5. Where are the top10 videos viewed from?
6. Time span during which the videos are mostly created, viewed, shared?
7. What are the top 10 video feeds by location, time?