Data Analytics & Machine Learning

Bootcamp Graduation Project
Nov. 2022

The Streaming Video Industry

According to PwC's latest Global Entertainment & Media Outlook 2022, the streaming video industry was valued at $79.1 billion in revenues worldwide in 2021 and will continue to grow at a pace of 7-10% annually for the next few years. The big driver of opportunity is a major shift by all major players in the subscription video space (Netflix, Amazon Prime Video, Disney+, Paramount+, Peacock) to hybrid streaming models that combine lower-priced, ad-supported tiers with more premium, ad-free tiers.

Across the streaming video industry, content providers and distributors are moving into big data to analyze subscriber funnels and viewership patterns to optimize content production/acquisition costs, help with programming decisions, improve content recommendation to their users and ultimately drive subscriber and advertising revenue.

Our Client

Our real-world client is an online distributor of online video based in Western Europe and specialized in music content, mostly full-length concerts and documentaries. They offer 3 advertising-funded, linear channels (free/no subscription), a premium subscription on-demand service and a large library of titles available for sale to other conventional channels.

Their channels are available worldwide and designed primarily for free, ad-funded platforms such as PlutoTV, Roku, YoutubeTV, Plex, Samsung TV Plus, LG Channels and other streaming and OEM services. Their premium SVOD service is available primarily through traditional pay-tv distributors and on a direct-to-consumer basis.

Each of the 3 channels (and each piece of content on the channels) are embedded with unique identifiers and markers 24/7 to automate and maximize advertising sales: position and duration of the ad breaks available, viewer profile, location by country, device, viewership history etc. All those parameters are used to customize the experience and offer as much audience targeting as possible to advertisers, who are then willing to spend more to reach a more targeted audience.

The Data

The client has made the following available to us:

- 550 CSV files containing viewership data for 3 channels across 14 operators in 67 countries from 18Feb21 to 6Nov22
- 1 CSV file containing advertising revenue for their 3 channels across 19 main territories from 28Feb22 to 25Oct22 (partial data to preserve some level of confidentiality)
- Programming details, program names, IDs, genres, keywords
- Channels mapping table
- Operators mapping table
- Countries mapping table

The data has been anonymized to preserve the confidentiality of the client.

Our Objectives

As the Client rolls out their services across an increasing number of platforms around the world and the number of viewers increases as well, the amount of data collected daily skyrockets exponentially. At this stage in their development (3 years after commercial launch), the Client collects on average 120,000 lines of viewership data per day and up to twice as much advertising data depending on granularity. The Client is no longer able to cope with traditional Excel and PowerBI tools, and is now looking to hire one or more data analysts to take the company to the next level.

Our study will focus on viewership, programming and advertising revenue data for the 3 linear channels to help answer the following:

- identify viewership patterns by channel, content, country, platform... ;
- identify revenue trends and determine which channel/content/genre brings in more revenue, by country or region ;
- make revenue projections into the next 2 fiscal quarters ;
- make content recommendation based on internal content tags.

EDA Methodology

Viewership Data

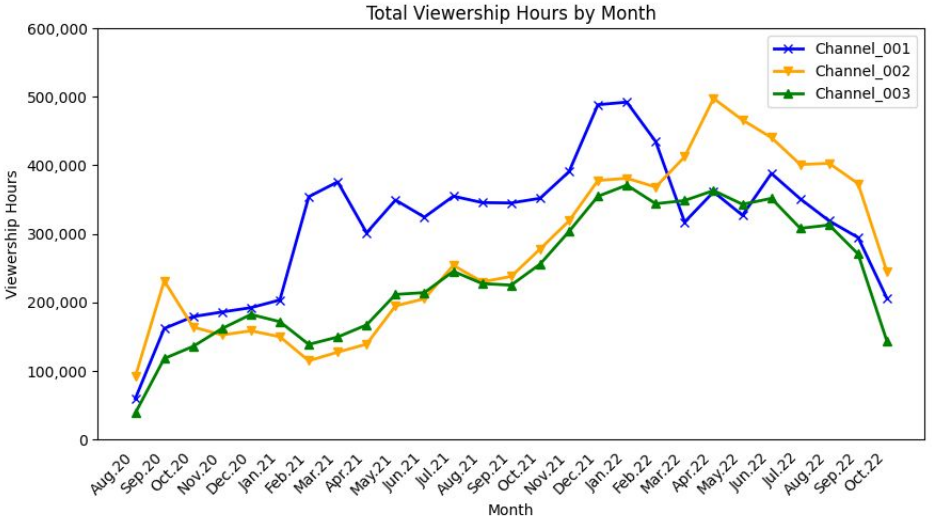
- Merge all 550 viewership files into one
- Anonymize the data pertaining to channel and operator names using anonymization key table. (CONFIDENTIAL DATA, NOT AVAILABLE TO PUBLIC)
    - Applied a function to each row to find string in list of anonymization key table for both channel and operator under unanonymized and now removed 'channel' column
    - Code should be refactored in future revisions or before deployment to improve speed. Currently takes 500 minutes to run for loop for each function.
- Anonymized 'content_id' column containing exact name of media content provided using Media Library key from data provider for programs and generated an anonymization key table for playlists. (CONFIDENTIAL DATA, NOT AVAILABLE TO PUBLIC)
    - Applied regex filters to obtain program or playlist numbers and created new columns containing anonymization key
    - Merged columns with .fillna
- Used Anonymized 'content_id' to match with genre data from data provider.
    - Merged columns on anonymized 'content_id' and removed extra columns.
    - Method was applied to 52 million rows with about 3000 rows of anonymized data keys. Method of using PANDAS .merge (similar to SQL or Excel VLOOKUP) is faster by about 250x
- Final DataFrame was exported as .csv for development purposes, but should be uploaded to PostgreSQL database.

Revenue Data

- drop columns as indicated by the client: "'bid_timeouts_rate", "render_rate", "fillrate", "avg_winning_bid (,Ç¨)" and "avg_imp_ecpm (,Ç¨)"
- convert date column from 'object' to 'date' with to_datetime
- drop rows that contain either all null values OR an "endpoint_request" value and all null values otherwise (853 rows)
- drop rows without a "country" value (12 rows)
- convert country codes to country names, and add a "region" column
- create new columns for CPM and pod drop rates
- replace "no viewership data" values in "channel" and "operator" columns with "unknown" (7,217 rows)

Our Initial Analysis

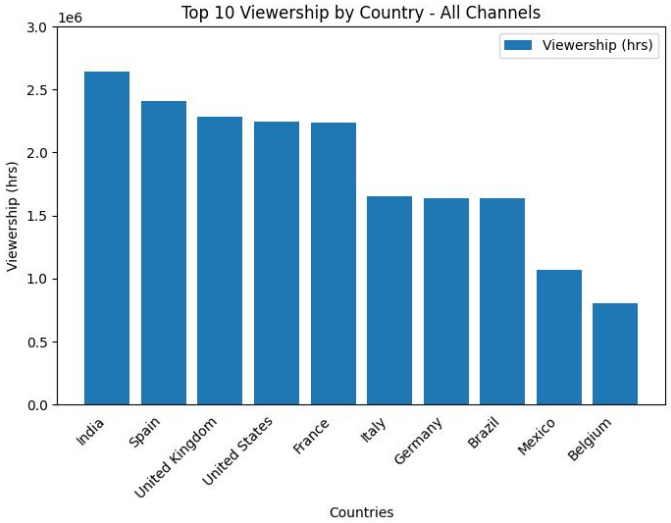
Total viewership increases five-fold until summer 2022, then dips dramatically



Strong growth overall for 2 years despite cyclical ups-and-downs, up five fold between Aug. 2020 and Q2 2022

Strong dip across all 3 channels from Jul. 2022

Content is popular all over the world



India comes out on top as the largest country in terms of consumption, all channels combined, followed by key Western markets.

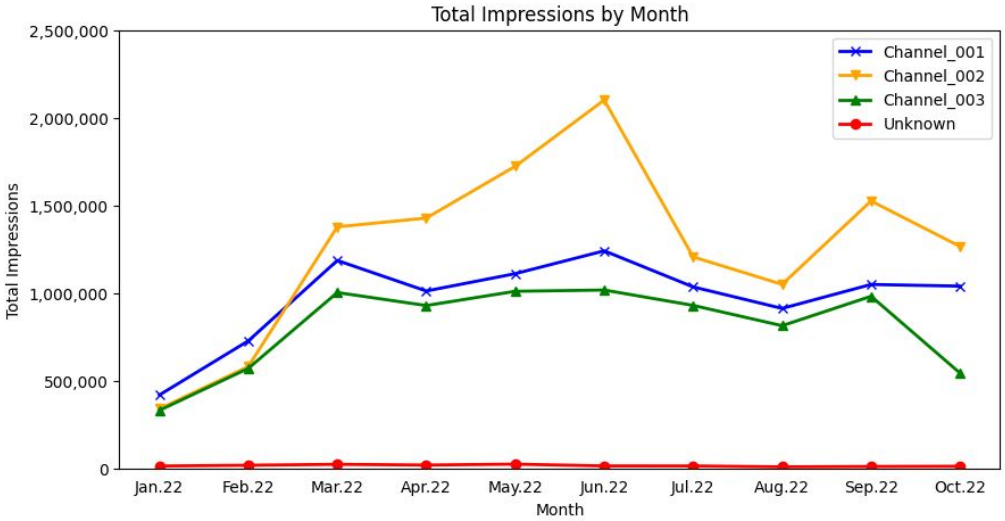
Brazil and Mexico are also in the top 10

Viewership rankings differ per channel

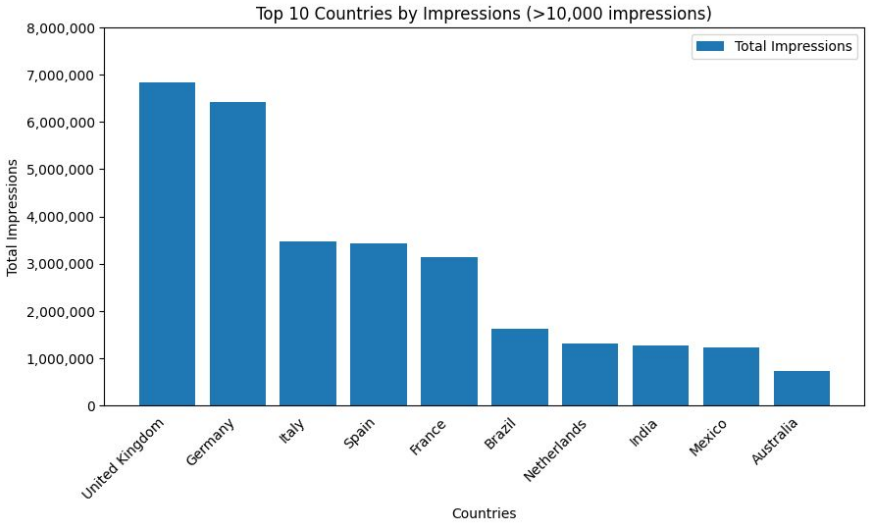
	Channel_001	Channel_002	Channel_003
1	Brazil	India	Spain
2	United States	United Kingdom	United States
3	France	France	United Kingdom
4	Spain	Spain	India
5	United Kingdom	Germany	France
6	Mexico	Brazil	Italy
7	Italy	Italy	Germany
8	India	United States	Australia
9	Germany	Mexico	Belgium
10	Belgium	Belgium	Netherlands

Rankings vary but the list of countries remains fairly consistent across the 3 channels, which could indicate a general interest for the product (music concerts offered on FAST basis) and simply different tastes in music.

Impressions follow revenue trends (unsurprisingly)

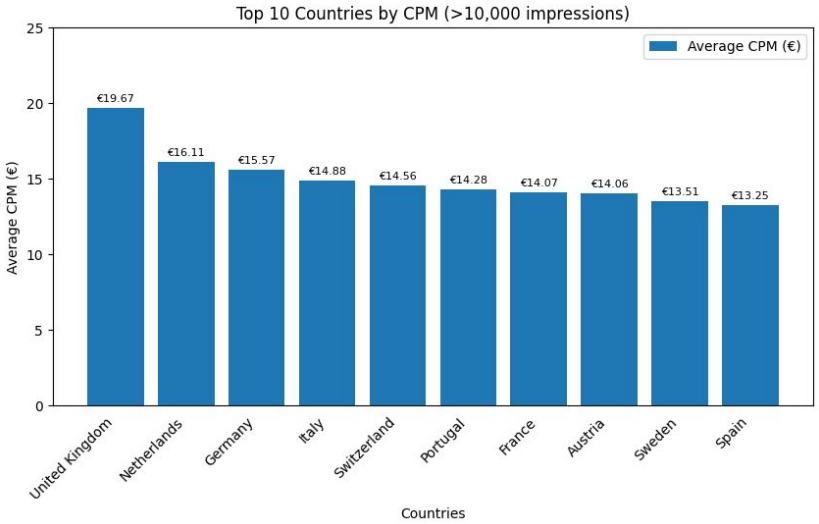


Western Europe delivers the most ad impressions



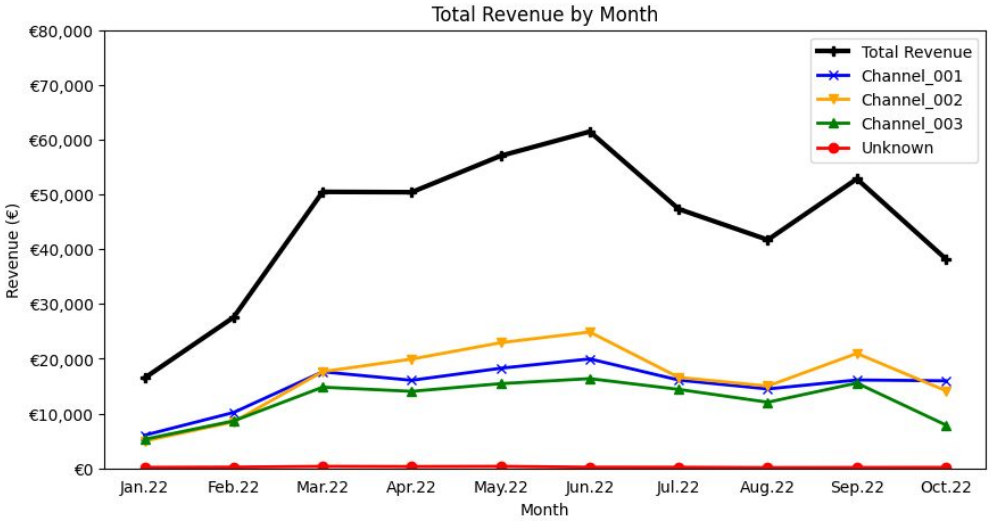
Brazil, India and Mexico deliver decent numbers of impressions, but CPMs are low in those territories so they eventually contribute little to overall revenue.

Highest CPMs are found in Western Europe



Surprisingly the US only delivers an average CPM of €12 (ranked #18), but our sample size here is small (only ~25k impressions)

All 3 channels' revenue trend in the same direction

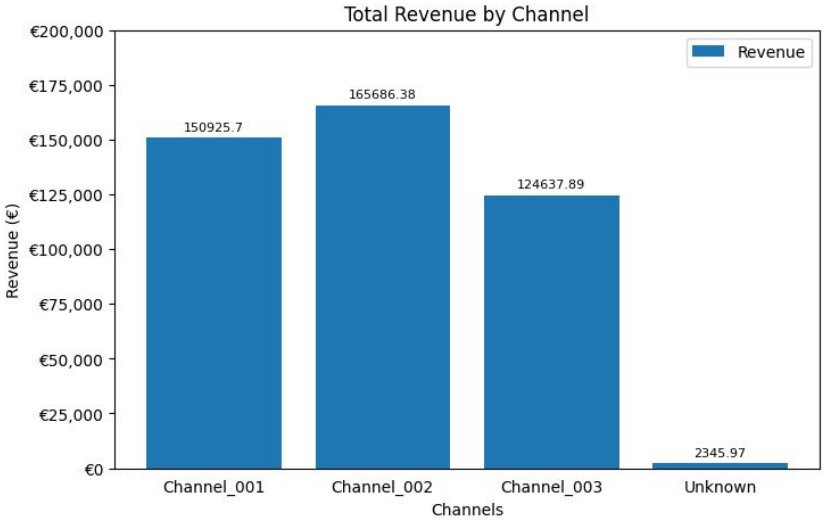


Channel_002 brings in slightly more revenue than the other two

"Unknown" is insignificant

Dip in summer months fairly typical (less viewership during European summer holidays), but the decrease in Sep-Oct is more concerning for ch. 2 and 3

Pretty balanced portfolio overall

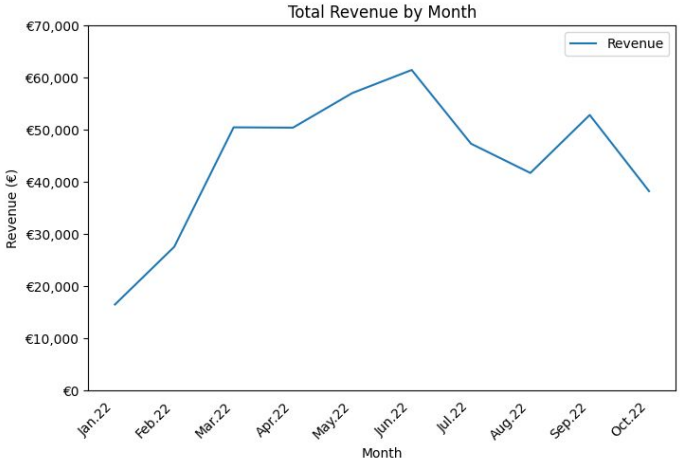


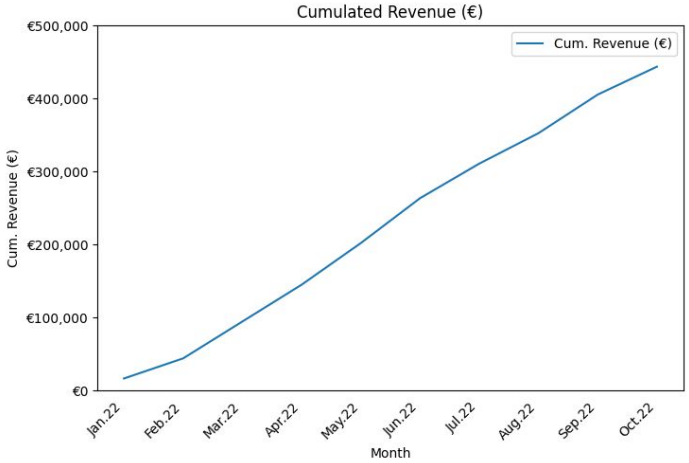
All 3 channels generate equivalent levels of revenue

Channel 3 brings in slightly less

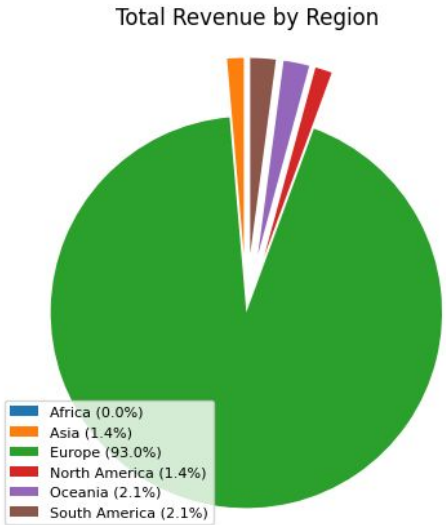
“Unknown” is insignificant

Overall healthy, growing revenue in YTD 2022 – although cyclical





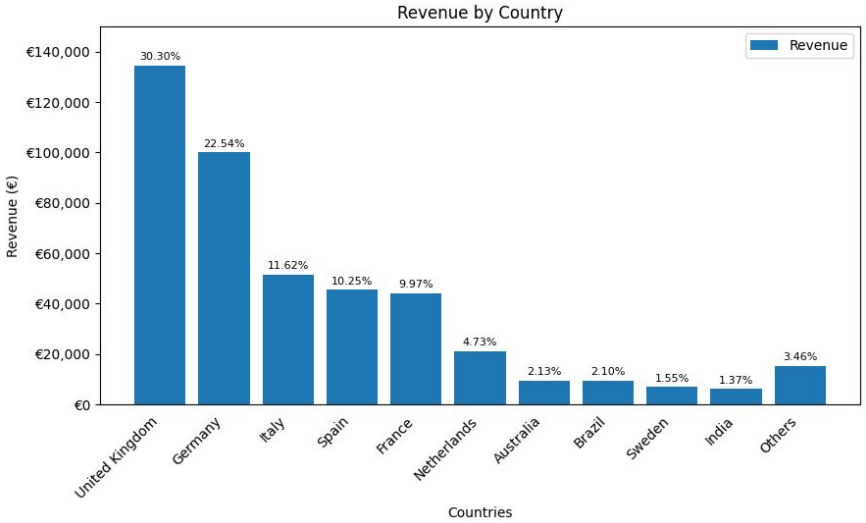
Revenue heavily reliant on Europe



Europe accounts for 93% of revenue in YTD 2022

Other regions are virtually insignificant

Top 10 countries generate 96.5% of all revenue

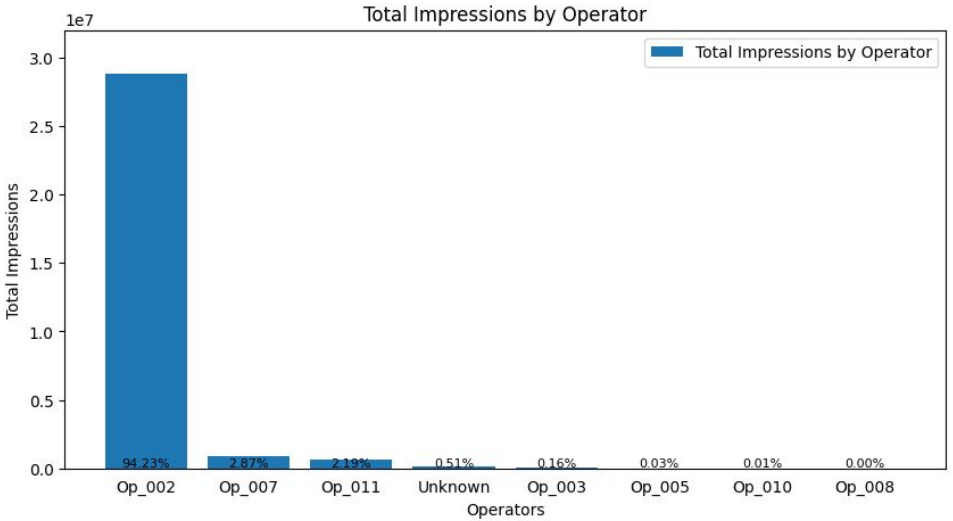


All Top 5 are in Western Europe

Top 5 generating countries total 85% of all revenue, and Top 10 countries generate 96.5% of all revenue

Portfolio heavily unbalanced, but understandable for a young, growth-stage start-up

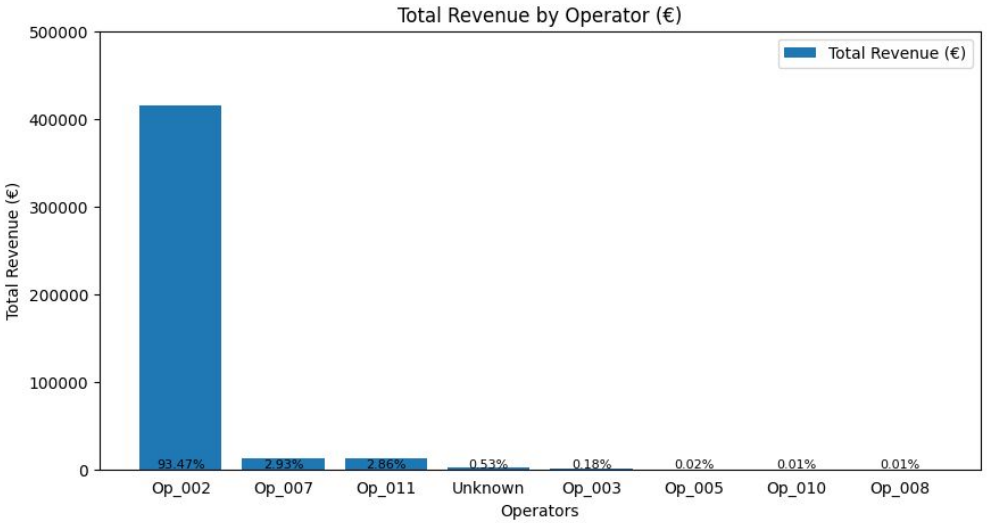
Advertising is massively dependent on one operator...



This is a huge strategic risk for the company to be so dependent on one client

94% of total impressions are delivered by Op. #2

… as is revenue



This is a huge strategic risk for the company to be so dependent on one client

93% of total revenue is generated by Op. #2

Initial Conclusions

Overall healthy, steadily growing revenue in YTD 2022

Top 5 countries in Western Europe deliver the most viewership, the highest number of impressions and the highest CPMs - therefore driving the most revenue

Revenue balanced pretty evenly across the 3 channels, but massively reliant on one operator and a handful of territories - HUGE STRATEGIC RISK

Company must try to diversify sources of revenue:

- Investigate why Op. 02 is so successful vs. others: penetration? territories? content more in line with viewers tastes? better channel exposure / marketing? etc.
- Explore why existing operators deliver so few impressions vs. Op. 02: detailed viewership/programming analysis, ad ops issues, ad sales strategy, SSP issues etc. ?
- Invest in business development strategies, launch with new partners/territories, invest in channel/content marketing to drive viewership etc.