Data Analytics & Machine Learning

Bootcamp Graduation Project
Dec. 8th, 2022
Jordan Bowman, Julien Dutronc, Ahmed Mossa, Peter Nguyen

The Streaming Video Industry

- The streaming video industry was valued at $79.1 billion in revenues worldwide in 2021
- Anticipated to grow 7-10% annually for the next few years.
- Driver of opportunity
    - Switch to hybrid streaming models that combine lower-priced, ad-supported tiers with more premium, ad-free tiers.
    - Utilizing big data in order to:
        - Optimize content production/acquisition costs
        - Help with programming decisions
        - Improve content recommendation to their users
        - Increase subscriber and advertising revenue.

Our Client

Our real-world client is

- Distributor of online videos specializing in music content, mostly full-length concerts and documentaries.
- Based in Western Europe
- 3 advertising-funded, linear channels
    - free/no subscription
    - premium subscription on-demand service
    - large library of titles available for sale to other conventional channels.

Each of channel collects data on:

- position and duration of the ad breaks available
- viewer profiles
- location by country
- device
- viewership history

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.

Premium SVOD service is available primarily through traditional pay-tv distributors and on a direct-to-consumer basis.

Our Objectives

The problem:

- Expansion and growth have led to large amount of data collection
  - Average of 120,000 lines of viewership data per day
  - Data on advertising has doubled over last three years
- Previously used Excel and PowerBI but with larger datasets more sophisticated tools are needed

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

- Explore viewership patterns by channel, content, country, platform, etc.
- Identify revenue trends and determine which channel/content/genre brings in more revenue, by country or region
- Make revenue projections

EDA Methodology

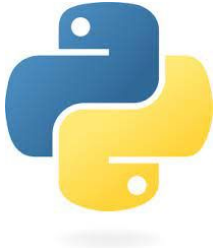
The Data

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.

Technologies, languages & tools used

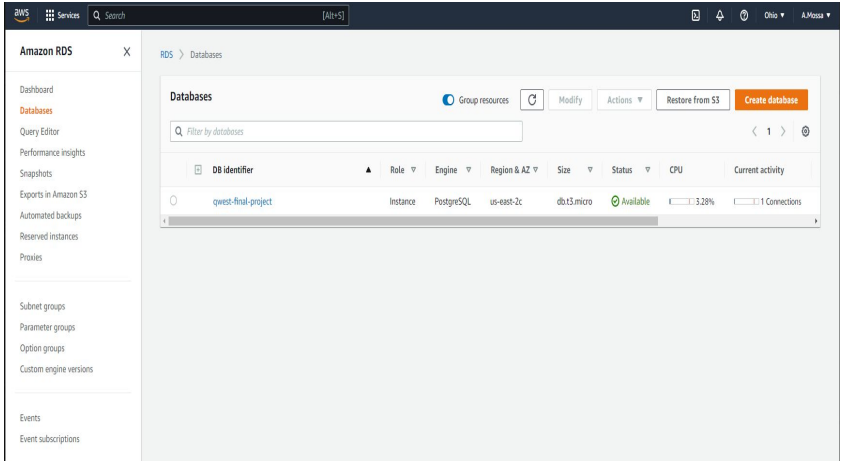


- Data Cleanup: Python 3.7.13, 3.9.13
- Exploratory Data Analysis: Python 3.7.13, 3.9.13, Microsoft Excel
  - Libraries Used: PANDAS, Matplotlib, numpy, sqlalchemy
- Machine Learning: Python 3.9.13
  - Libraries Used:
    - MLV1: Python 3.9.13, PANDAS, PySpark, Matplotlib, Numpy, sklearn.linear_model Linear Regression
    - MLV2: Python 3.9.13, PANDAS, PySpark, Matplotlib, Tensorflow_gpu, Keras, sklearn.model_selection train_test_split, sklearn.preprocessing StandardScaler and OneHotEncoder
    - MLV3: Python 3.9.13, PANDAS, PySpark, Matplotlib, Tensorflow_gpu, Keras, sklearn.model_selection train_test_split, sklearn.preprocessing StandardScaler and OneHotEncoder
    - MLV4: Python 3.9.13, PANDAS, PySpark, Matplotlib, statsmodels.tsa.statstools adfuller, statsmodels.api, itertools

Database Choice

- The team decided on using a live database on AWS utilizing the RDS service.

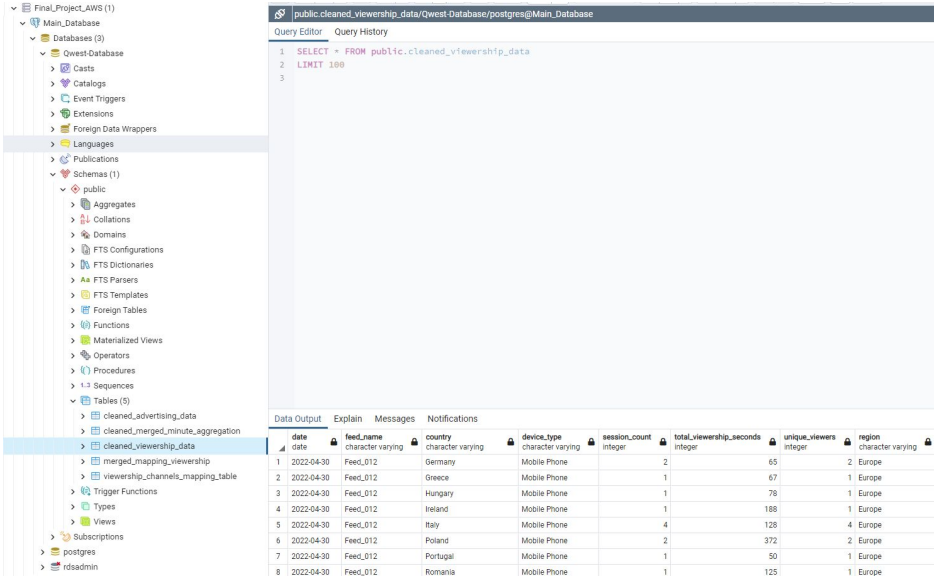
- A database was created to host the cleaned tables we have for the analysis, including (cleaned Viewership data, cleaned Advertising data, and cleaned minute aggregation data).



Database access

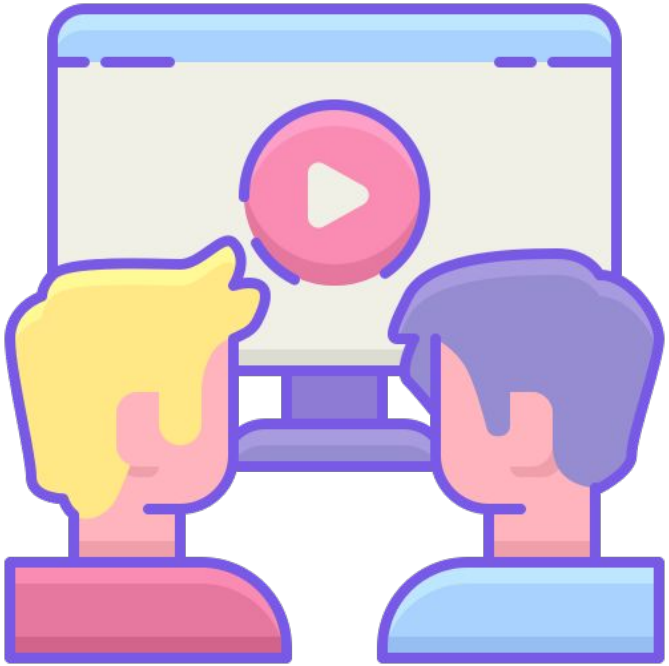
Any further cleaning or manipulation of the database was done in PgAdmin using postgresql.

The database was connected also to the dashboard, so every interaction is done by querying the database for updated results in case the data changed.



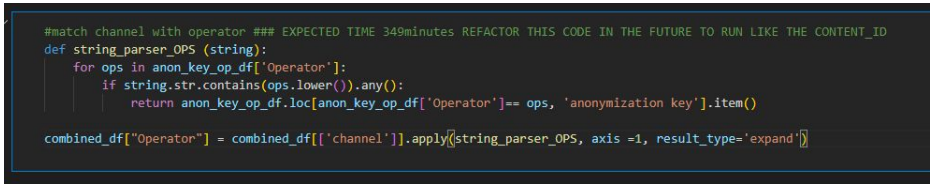
Viewership Data

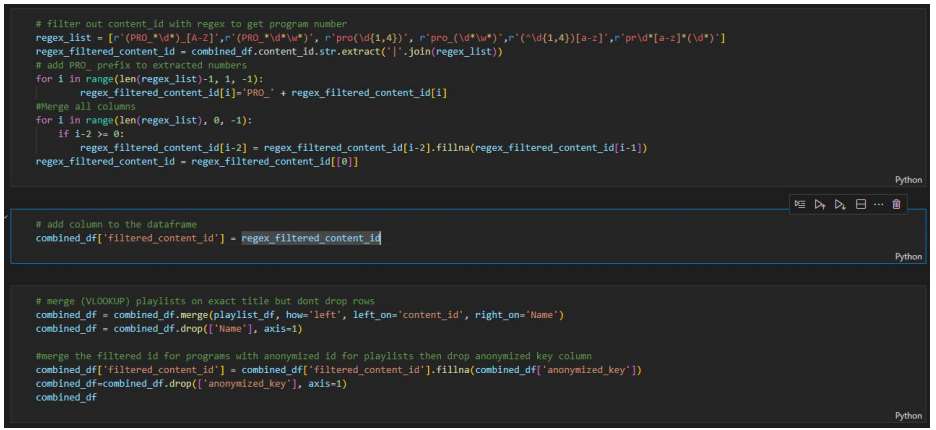
- Merge all 550 viewership files into one
- Anonymize the data pertaining to channel and operator names using anonymization key table.
    - Applied a function to each row to find string in list of anonymization key table for both channel and operator under unanonymized and now removed 'channel' column
- 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.
    - Applied regex filters to obtain program or playlist numbers and created new columns containing anonymization key
    - Merged columns with .fillna



Viewership Data

- 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
- Final DataFrame was exported as .csv for development purposes.



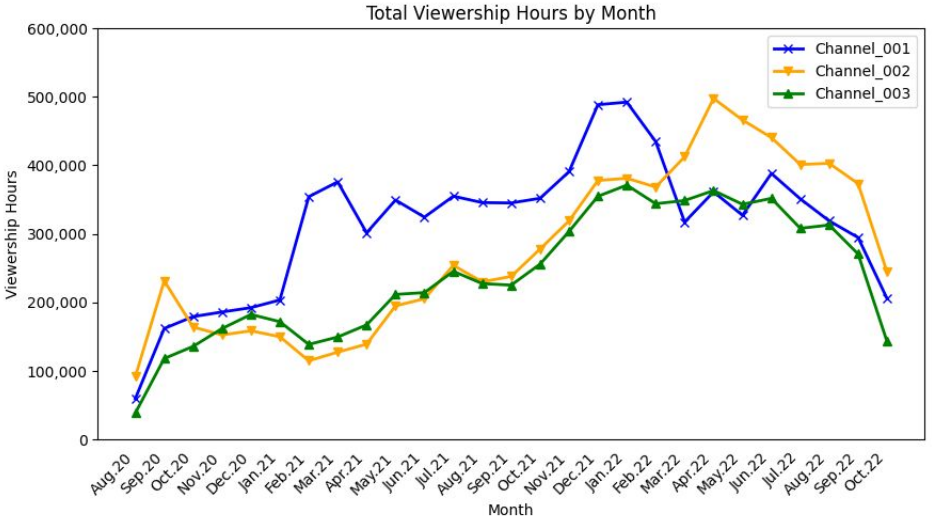


Revenue Data

- 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

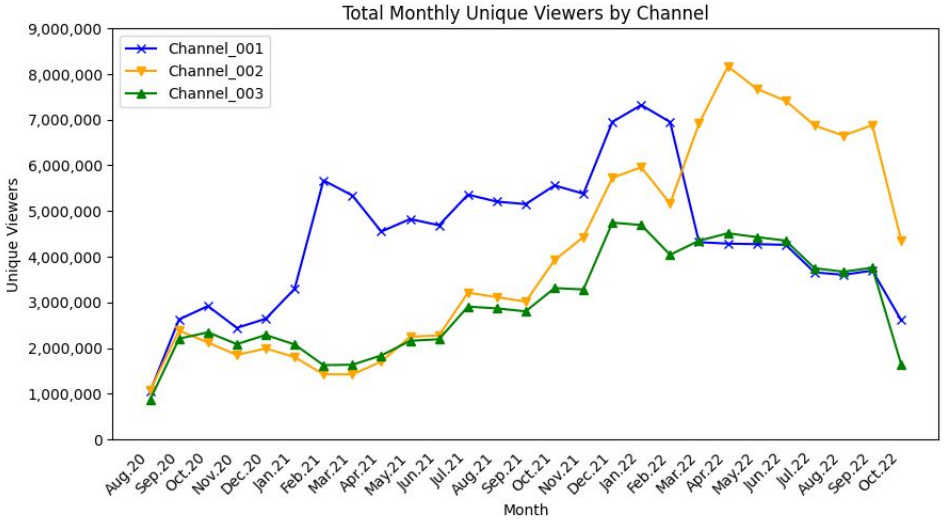
Total viewership increases five-fold until summer 2022, then dips dramatically



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

Unique viewers follow the same trend

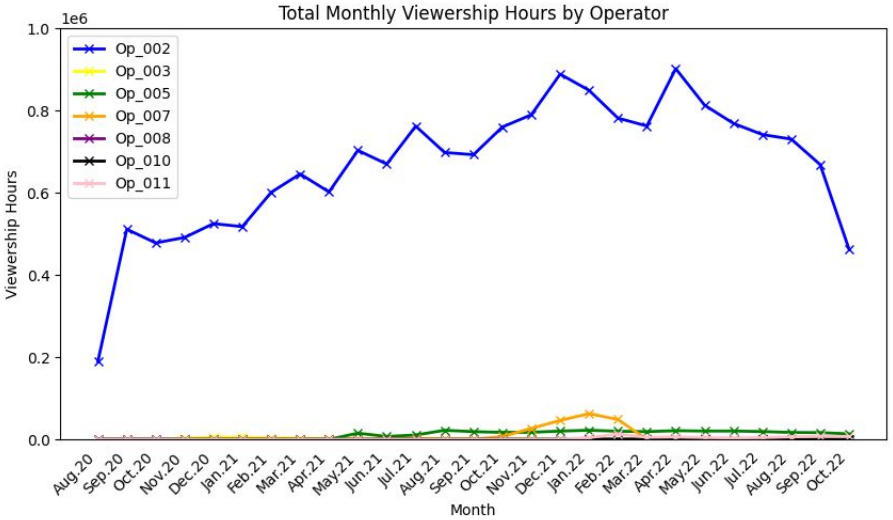


Number of unique viewers grew consistently until Q2 2022, then started dipping quite dramatically

Ch_001 dipped first, early in 2022

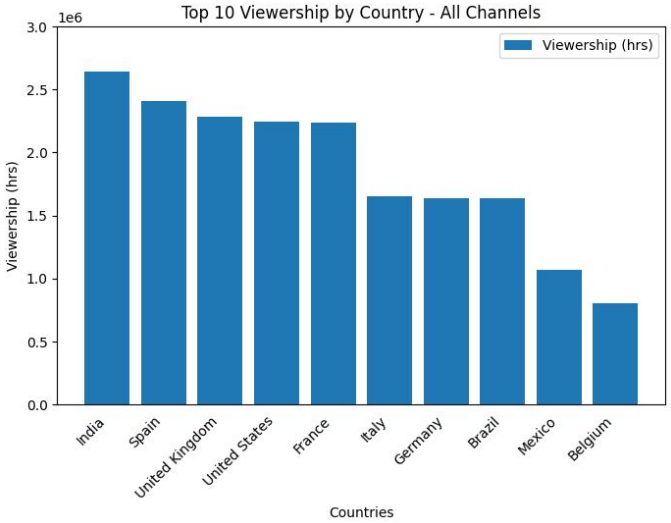
All 3 channels dropped in the same order of magnitude in Sep. 2022

Op. #2 generated 81% of total viewing hours over the period



Being so reliant on one operator to generate viewership hours (and therefore generate advertising revenue) is a significant strategic liability for the company.

Content is popular all over the world



India is the largest country in terms of consumption, all channels combined, followed by key Western markets.

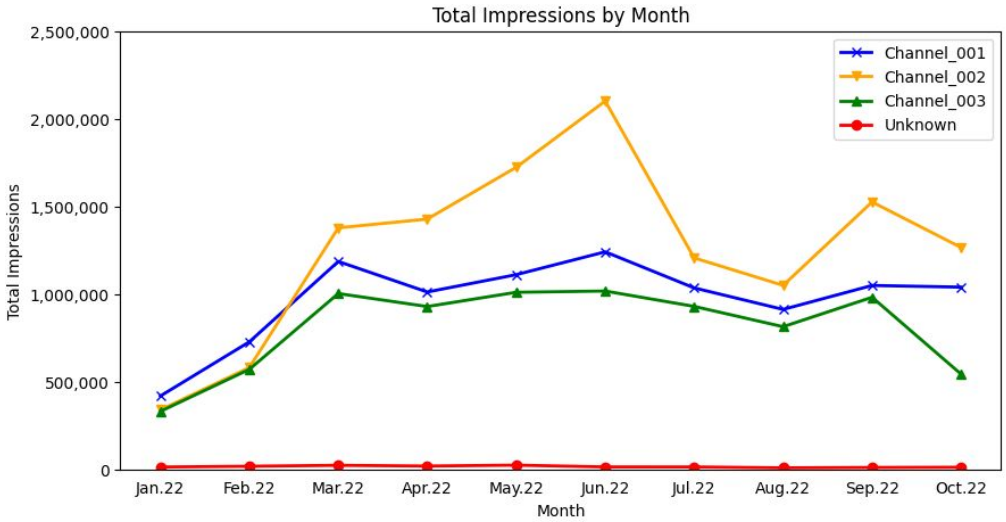
Brazil and Mexico are also in the top 10

Viewership rankings differ per channel

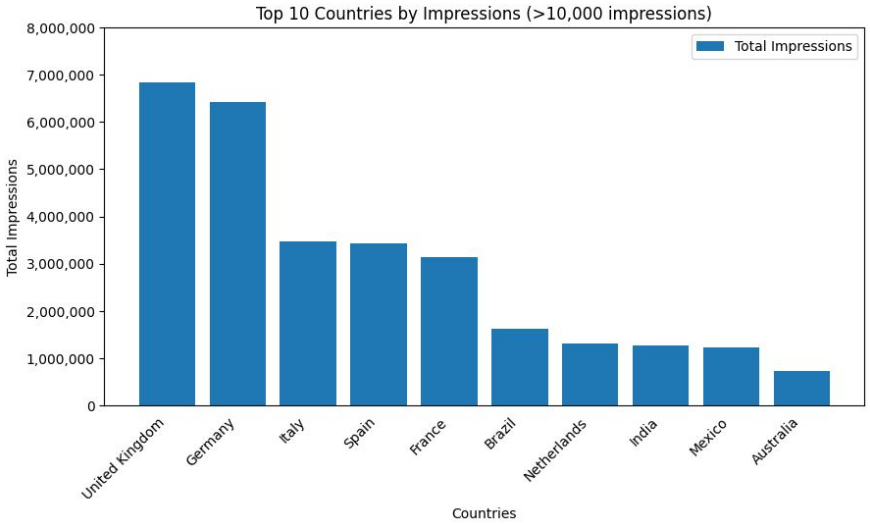
	Channel_001	Channel_002	Channel_003
1	Brazil	India	Spain
2	United States	United Kingdom	United States
3	France	France	United Kingdom
4	Spain	Spain	India
5	United Kingdom	Germany	France
6	Mexico	Brazil	Italy
7	Italy	Italy	Germany
8	India	United States	Australia
9	Germany	Mexico	Belgium
10	Belgium	Belgium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

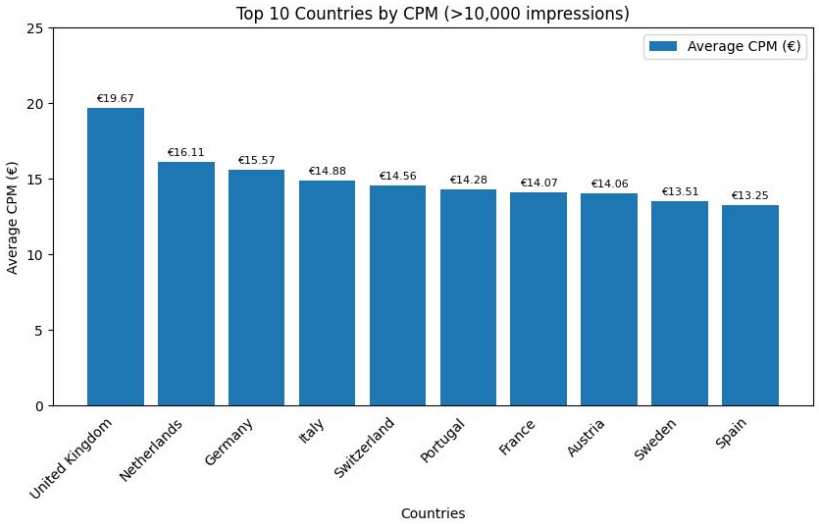


Western Europe delivers the most ad impressions



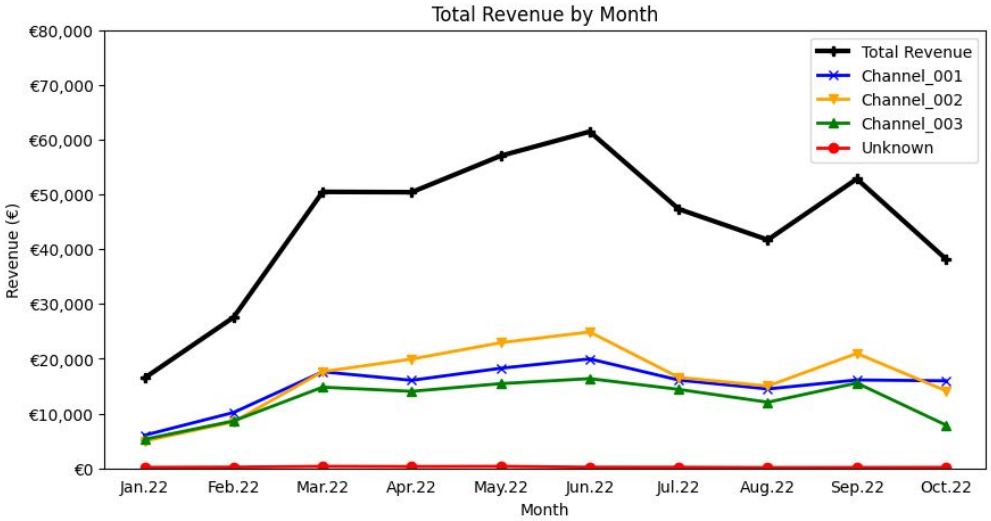
Brazil, India and Mexico deliver decent numbers of impressions, but CPMs are low in those territories so they eventually contribute little to overall revenue.

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

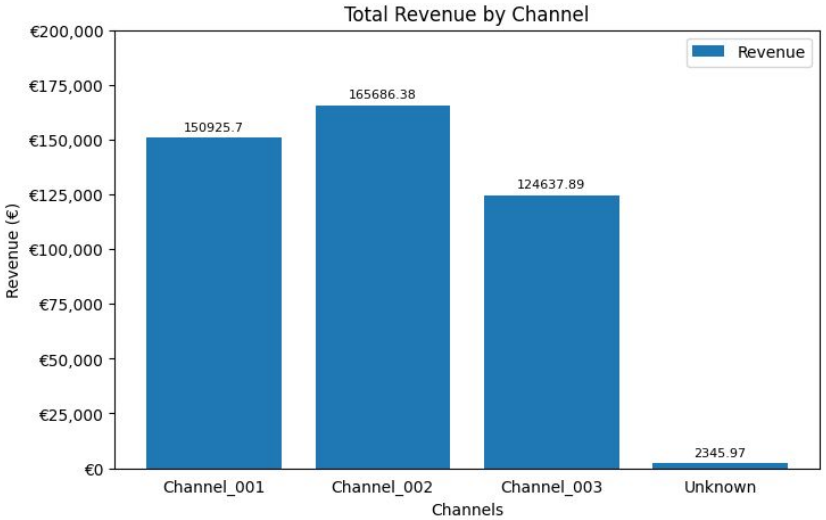


Channel_002 brings in slightly more revenue than the other two

"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

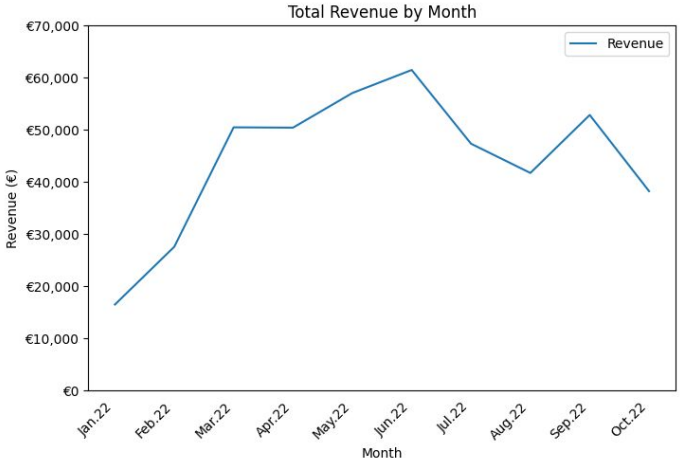


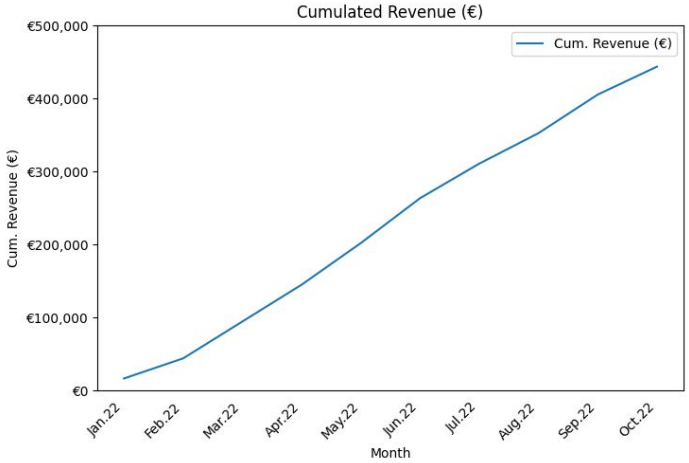
All 3 channels generate equivalent levels of revenue

Channel 3 brings in slightly less

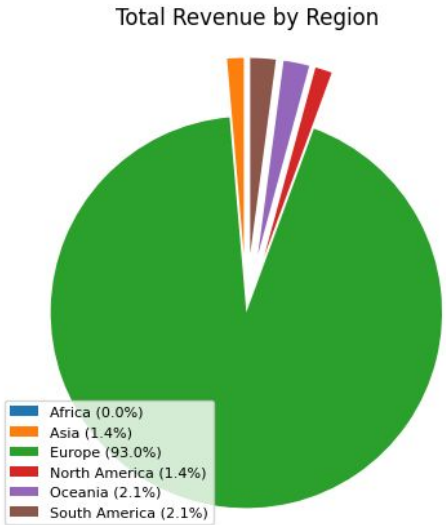
“Unknown” is insignificant

Overall healthy, growing revenue in YTD 2022 – although cyclical





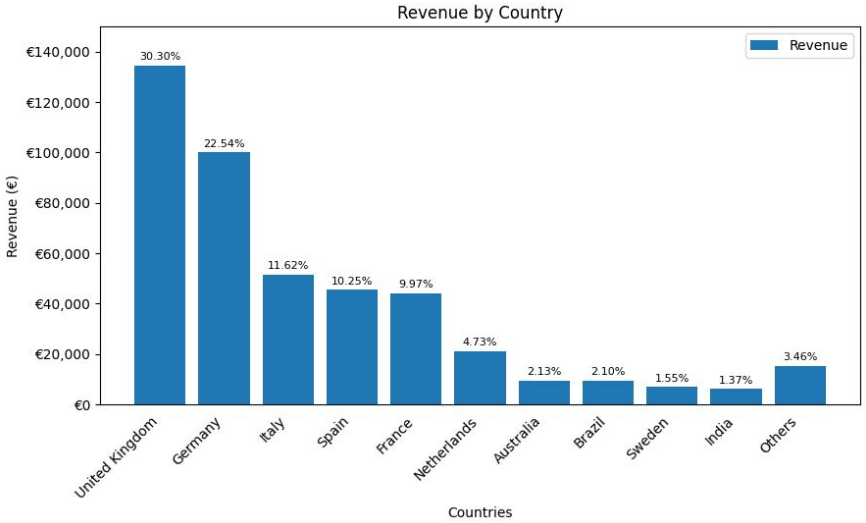
Revenue heavily reliant on Europe



Europe accounts for 93% of revenue in YTD 2022

Other regions are virtually insignificant

Top 10 countries generate 96.5% of all revenue

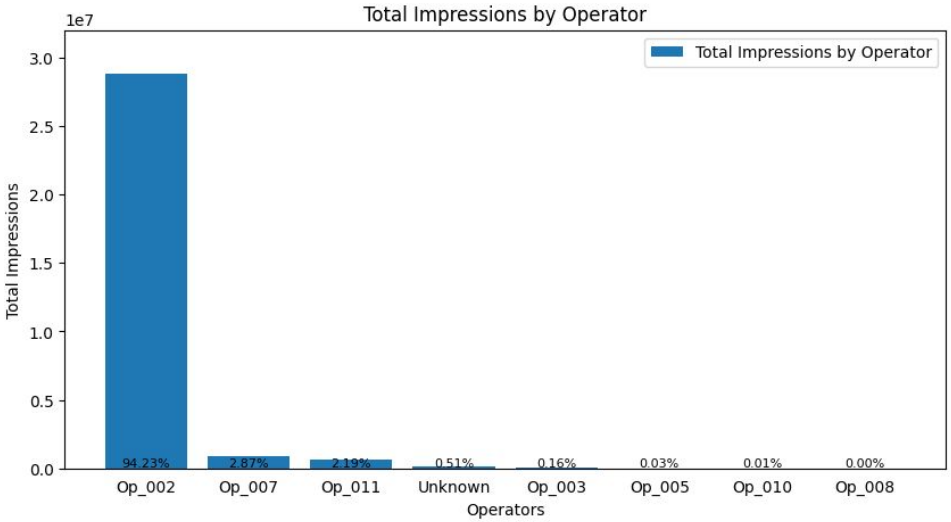


All Top 5 are in Western Europe

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

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

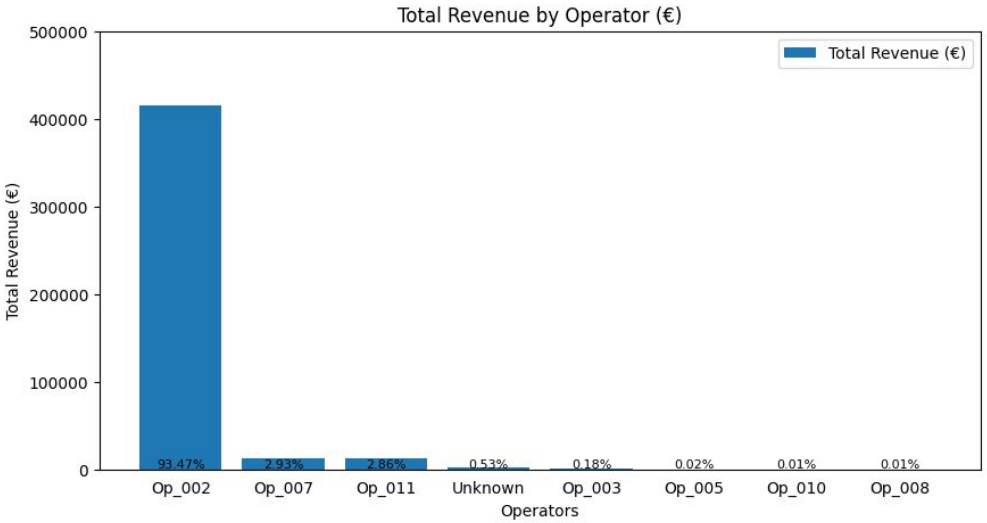
Advertising is massively dependent on one operator...



This is a significant 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 significant 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.

Dashboard

Tableau Dashboard

Focus on the 4 major KPIs:

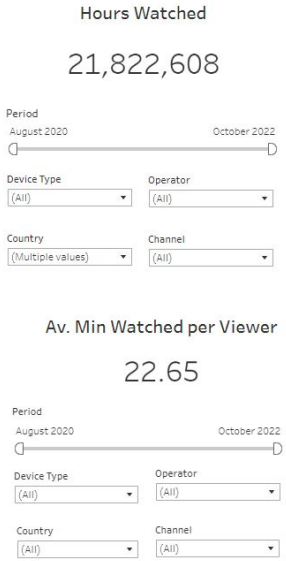
- Viewership: number of viewers and hours watched
- Revenue: ad impressions delivered and revenue generated

Questions we are trying to answer:

- Which countries/regions attract the most viewers and produce the most viewing hours?
- Which countries/regions generate the most advertising impressions and revenue?
- And more importantly, which countries/regions are the most attractive on a per-viewer basis? i.e. most hours per viewer and most revenue per viewer?

Tableau Dashboard

The live dashboard is available here: [Streaming Video Analytics | Tableau Public]



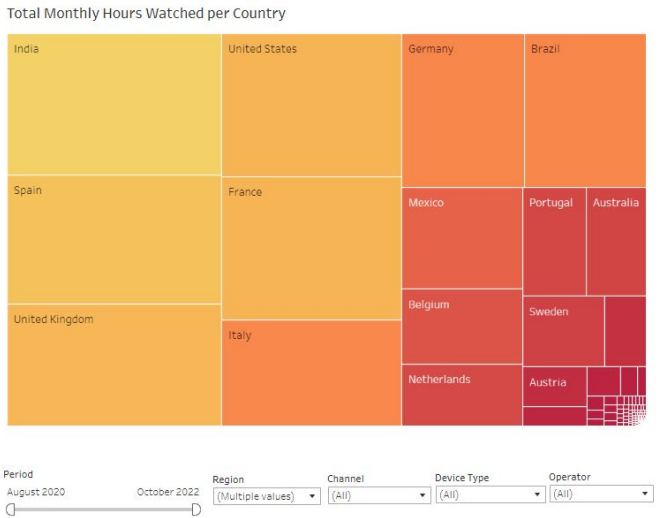
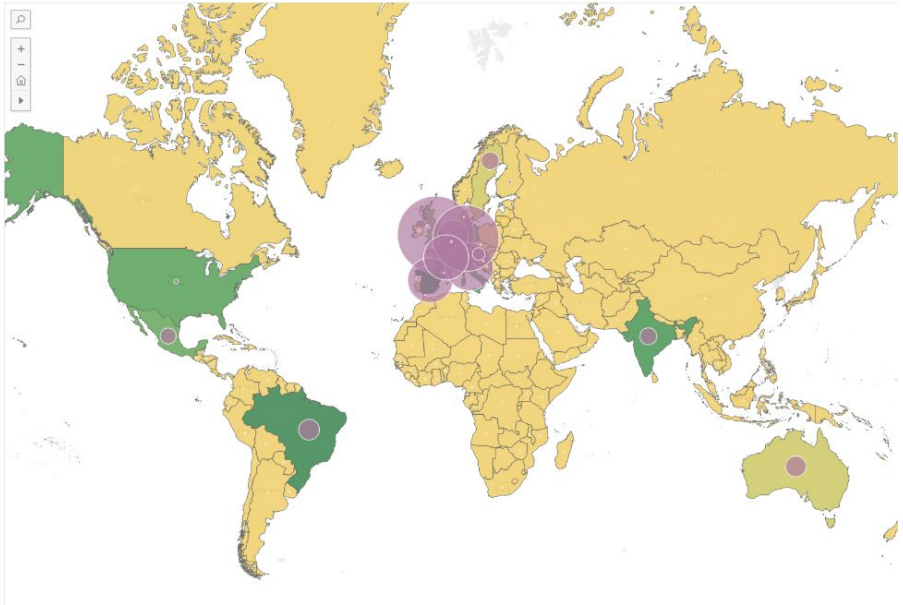


Tableau Dashboard

The live dashboard is available here: [Streaming Video Analytics | Tableau Public]

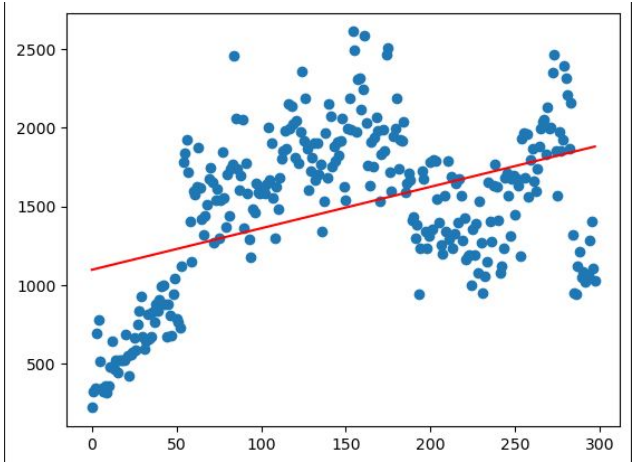


Machine Learning

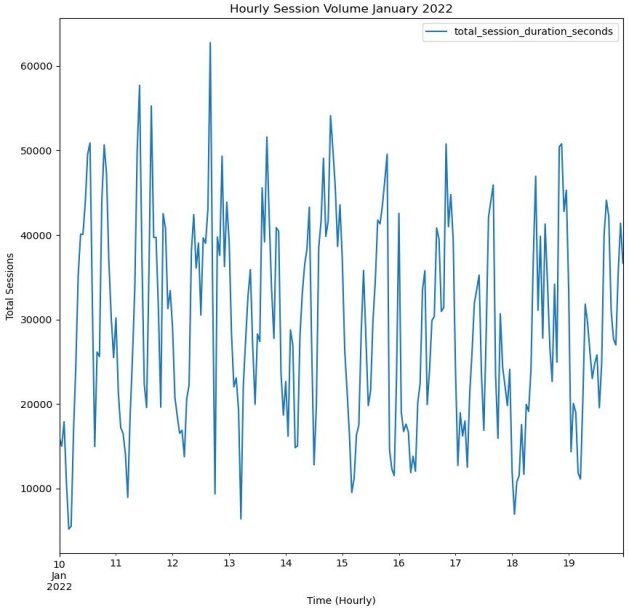
Data Exploration

Looking for trends and patterns in the data to decide what type of model to use.

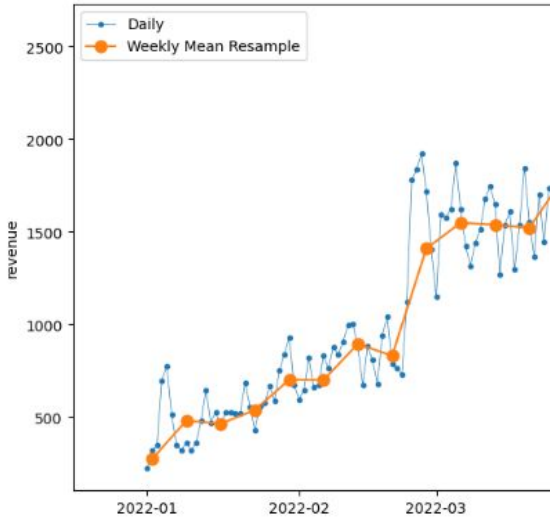
Linear?



Timeseries?

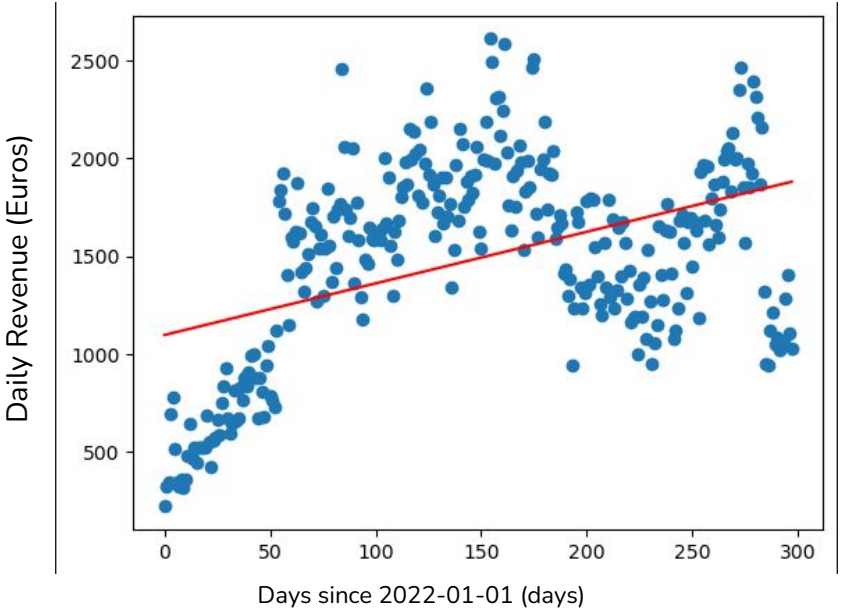


Moving Average?

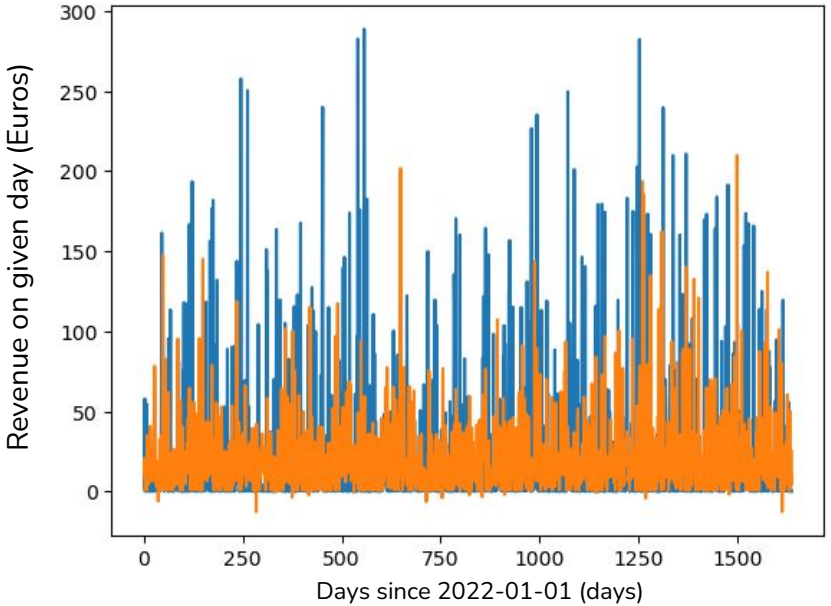


Four Different Models

MLV1: A Simple linear regression model of the revenue. This model was simple and fast to run, but inaccurate, did not fully utilize the available data, and was sensitive to outliers



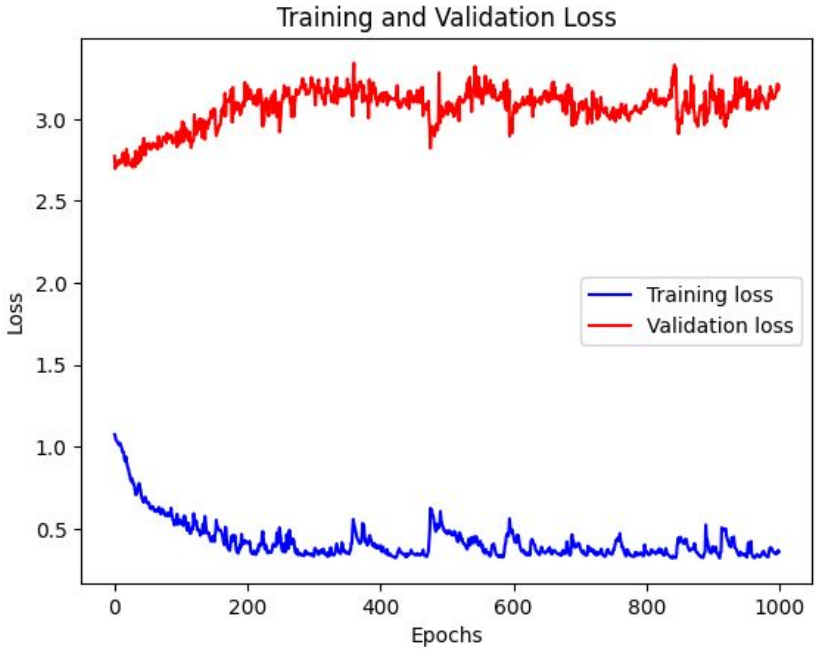
MLV2: Attempted to add additional data to the analysis by adding information from advertising data that could be known prior to airing program. Got better prediction, but was not elegant and required too much additional input.



Four Different Models

MLV3:

- Attempted to use a larger viewership dataset containing number of views at minute level aggregation to predict viewership.
- Accurate prediction of viewership could be later turned into a revenue prediction.
- Data was noisy due to having too many unique genre labels and caused overfitting.
- Significantly more data cleaning was required.



Four Different Models

MLV4: Attempted to build off of MLV1 and MLV3. Simplifying the model by removing categorical data all together, but keeping timeseries and applying non-linear regression. Utilizes Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model.

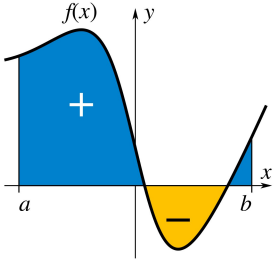
Seasonality

Look for seasonal trends

Auto-Regressive

Use lagging values to predict future values

Integrated



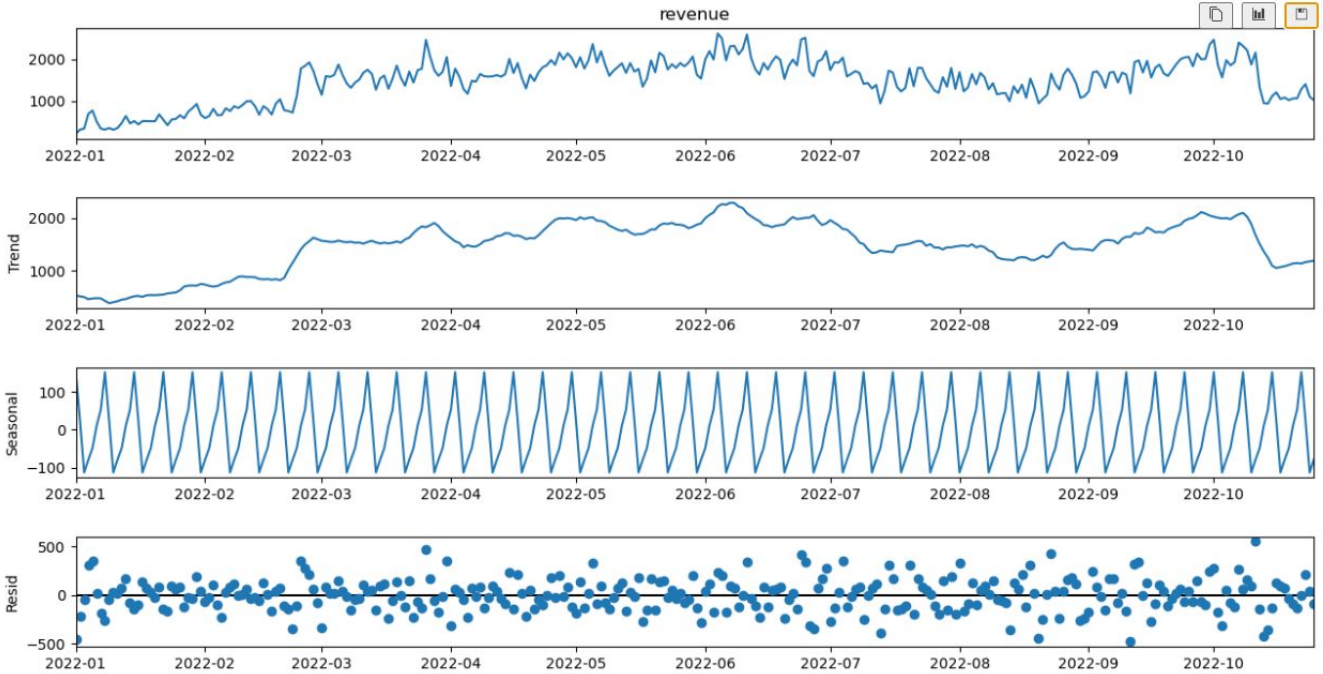
Instead of just using the raw value, also look at the differences between each point along timescale

Moving Average

Take the lagging prediction errors as inputs

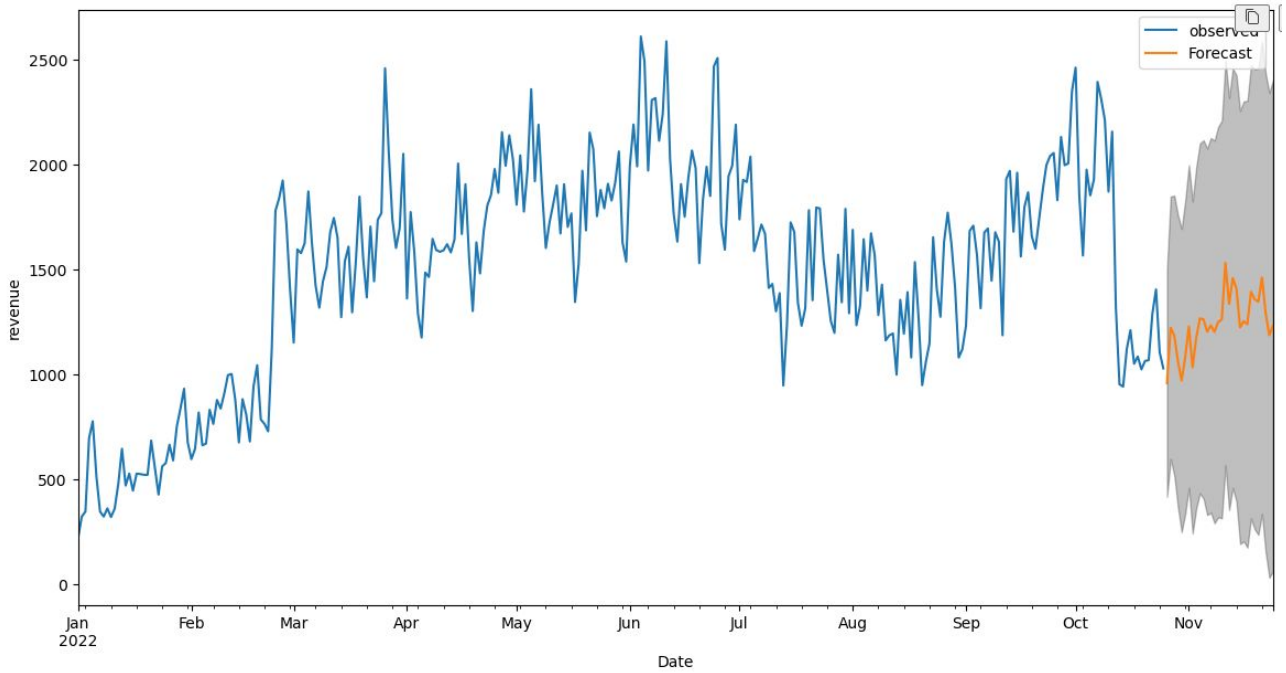
Results

Break it down, apply a Fourier Wave Transform. What components make up our revenue plot?



Results

Apply the SARIMA model, and predict the revenue!

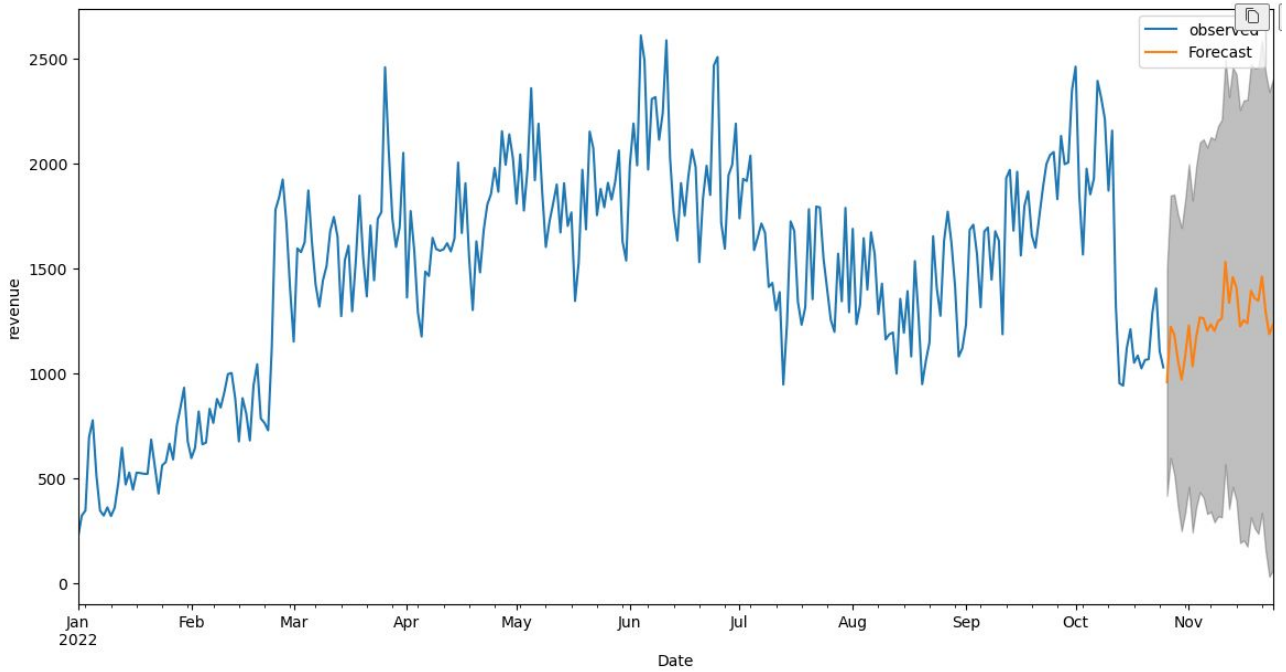


30 Day Forecast

	lower revenue	upper revenue
2022-10-26	417.03210102	1498.10324704
2022-10-27	596.69795242	1846.35530911
2022-10-28	518.28284252	1850.96277643
2022-10-29	365.33044904	1757.86063263
2022-10-30	248.15320550	1692.24574079
2022-10-31	336.28946223	1828.19164290
2022-11-01	458.88969804	1996.42496812
2022-11-02	242.10417236	1823.70918592
2022-11-03	360.60621224	1984.99464967
2022-11-04	433.15088233	2099.18796785
2022-11-05	408.06781527	2114.71719760
2022-11-06	329.39477086	2075.68783213
2022-11-07	340.49940903	2125.50400611
2022-11-08	290.42560654	2113.20048634
2022-11-09	319.05967222	2178.53656312
2022-11-10	314.10155285	2208.22434732
2022-11-11	565.67170718	2494.94550316
2022-11-12	352.93107632	2316.63777777
2022-11-13	458.86635776	2456.41077139
2022-11-14	393.75147714	2424.60243784
2022-11-15	191.54900298	2255.19694833
2022-11-16	204.13652015	2300.08638573
2022-11-17	174.56059268	2302.33286832
2022-11-18	313.21467939	2472.34673218
2022-11-19	261.73762641	2451.78415104
2022-11-20	235.48311654	2456.01592333
2022-11-21	335.48925884	2586.09669390
2022-11-22	152.11721954	2432.40341102
2022-11-23	31.35724422	2340.94128476
2022-11-24	61.76521242	2400.28033703

Results

Apply the SARIMA model, and predict the revenue!



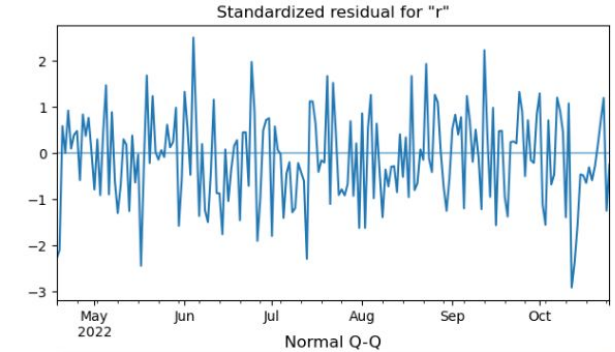
5 Day Forecast

	Date	Predicted_Mean	Lower Bound	Upper Bound
0	2022-10-26	957.56767403	417.03210102	1498.10324704
1	2022-10-27	1221.52663077	596.69795242	1846.35530911
2	2022-10-28	1184.62280947	518.28284252	1850.96277643
3	2022-10-29	1061.59554084	365.33044904	1757.86063263
4	2022-10-30	970.19947314	248.15320550	1692.24574079

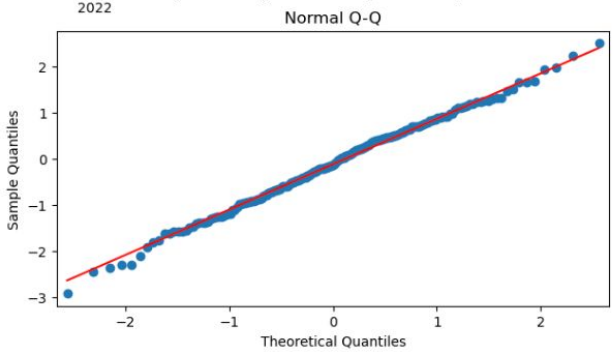
Results

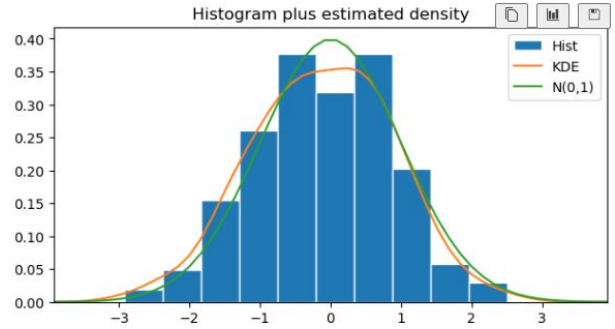
Model Statistics

How big is the difference between our prediction and actual?

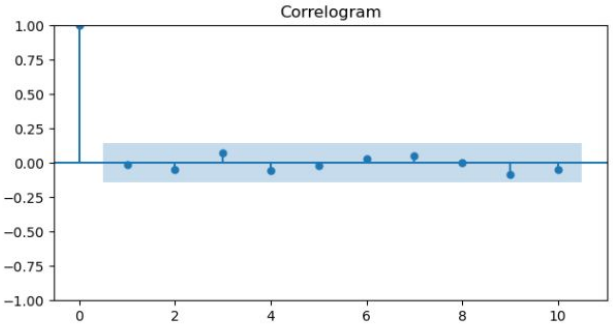


How normally is our data distributed?





After Detrending Data, how is our data distributed?

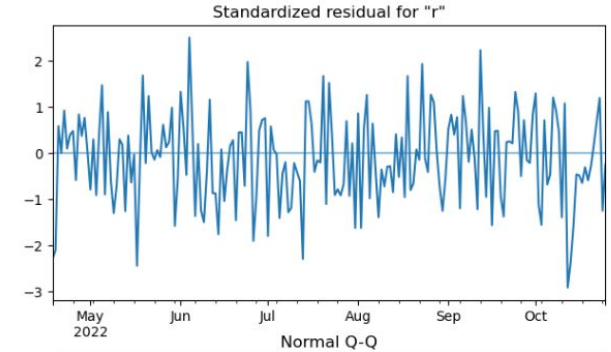


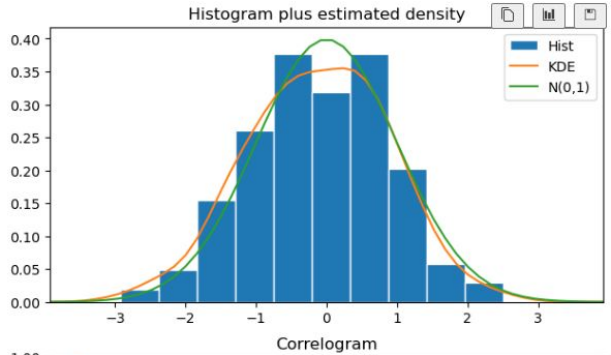
How random is our dataset at different time lag separations?

Results

Model Statistics

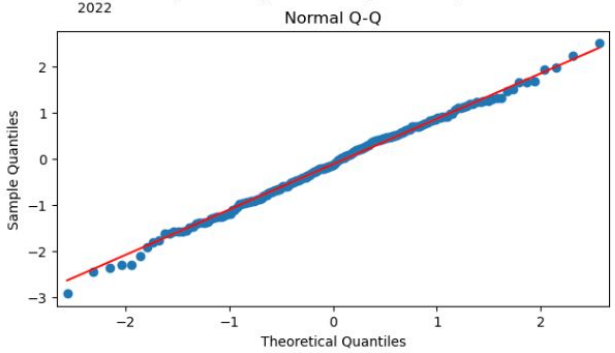
Some volatility seen, occasionally has significantly higher than or less than expected values

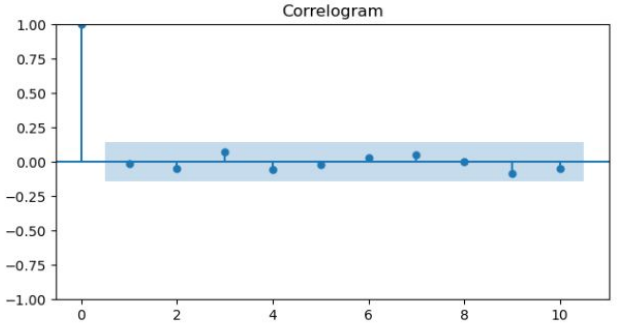




Again, normal distribution

Data is normally distributed, even at far ends





Aside from day-0, it is quite random.

Recommendations

Recommendations

- We would recommend to the company, creating and maintaining a unified data pipeline to ensure data integrity and uniformity.
- Diversifying the program licensing and not rely on one single provider for the majority of broadcast.
- Inject resources behind promising markets, and capitalizing on traction gained in regions such as: the United States, Australia, and South America.

What could be done differently

- Continue cleaning up data by binning genres for MLV3 and try to reduce noise to improve model
- What if we used ARIMA instead of SARIMA for MLV4?
- How can we utilize pod fill rate?

Questions