Comprehensive Analysis on Consumption of Entertainment Sources

Aditya Kumar Singh
(Data Analytics)
National College of Ireland
Dublin, Ireland
x20140410@student.ncirl.ie

Aryan Rajput
(Data Analytics)
National College of Ireland
Dublin, Ireland
x20128088@student.ncirl.ie

Javed Mohammed
(Data Analytics)
National College of Ireland
Dublin, Ireland
x19219458@student.ncirl.ie

Priyanka Ashok Sujgure
(Data Analytics)
National College of Ireland
Dublin, Ireland
x20136706@student.ncirl.ie

***Abstract*—Human entertainment is constantly evolving phenomenon and carries social and financial implications. The research analyzes selected four forms of entertainment namely Movies, Music, Books and Video-games via Data Analysis algorithms and elucidate trends and changes with respect to each other as well as over a period. The research also navigates through various genres popularity among each entertainment source and follow any specific change among popular choice across time. One of the important elements of the study is the conclusion extracted by the knowing that wide scale penetration of internet across common household has allowed for entertainment to be accessed remotely and at a faster pace. Streaming services has opened a new gateway for movies and music. Video-gaming has seen tremendous growth and in turn benefited the most from online campaign by developers. Print media has a strong foothold and is predicted to shift toward digitization in the near future.**

***K*eywords – Entertainment, music, movies, books, video-games, internet, genres.**

I. Introduction

"The most precious commodity possessed by a wise man is free time" by German playwright Paul Ernst appears in his "Diary of a Poet" [2]. Humankind has found countless ways to enjoy themselves in their spare time since the dawn of humanity, and these leisure practices have developed and flourished into diverse realms over time. The introduction of new media has also brought upon a drastic change in the fundamentals of the way people spend their time. It has resulted in various new business opportunities to be setup as well as formation of an organized sector. It is of utmost importance to study these trends closely and monitor population activity.Inline to the survey of GfK Media, Edition 05/2017 as shown in figure 1 the project has considered the top categories of entertainment for the general population as the entertainment market is ever increasing and is a great source of revenue [2]. So Music, Movies, Video games and Books was chosen as the study areas for analysis. The print media has been present here for the longest time. It is important to analyze the effects of newer competition on the overall. There is also a need to comprehend the effects of digital books or e-Books. Music has transformed from live club and street performances to recording studios producing songs at never-before-seen increased rate. Several apps and website provide live streaming services. Music has become fast paced and represent the changing culture of frequent kicks and rocks. Video games are the most recent phenomenon and has captured a new spot in the market. Computer has reached every household which in turn has made entertainment via computer easily accessible to all. The development of video-games took up post 2000 and have not seen any decline since. The movies industry has been going strong since 1950's. The setting up film academies has produced a new set of professionals which are thorough in film making. The setting up of cinemas has also made it possible to reach wider audience. It is also important to note the effects of internet on the cinema industry. All these fields carry importance in research as these influence the general trends globally.

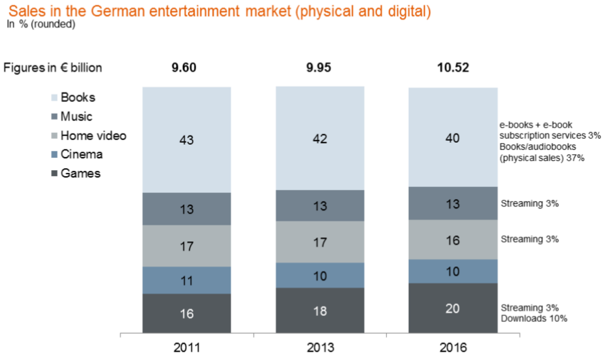

Fig. 1. Market Share for Entertainment

Media has long been regarded as a cultural influencer, with the entertainment industry at the forefront. With the introduction of the internet, the playing landscape has shifted, and conventional forms of content, such as print media, have seen increased competition. The music and movie industries have undergone significant expansion and investment. Millions of people around the world now have instant access to limitless resources thanks to the Internet which can be seen in figure 2. The video game industry though new as comparatively has highest per net growth. The online gaming has taken roots in the new generations and people still enjoy single player campaign more than ever. Since these developments have a strong financial impact, it is important to investigate them.

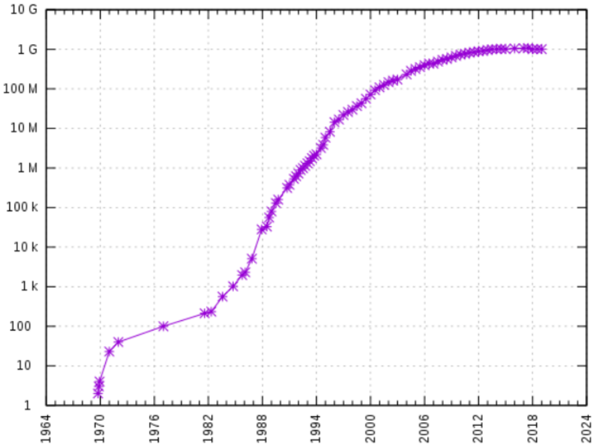

Fig. 2. Global Internet Growth

The research question that needs special mention is:

- Analyzing trends in each specific entertainment domain.
- Comparative growth study and visualization in the same frame of time among different domains.

II. Related Work

In [4], the author supplements the findings of behavioral changes linked to gaming namely violence and its addictive nature that the gamer experience. The latter has a positive correlation with the growth in the gaming industry that can be seen through analysis in the report. The study emphasized on the feature video-games showcases that binds the user and hence establish itself as viable source of entertainment.

[5] is a full comprehensive collection of all the information about video-game. The book discusses in depth about the growth of video-gaming from a side hobby to a full-on financial sector. The growth is attributed to the conversion of former gamers to developers and publishers.

The [6] is a book which gives an opposing opinion concerning the general notion about the adverse effect of internet on the music industry. The internet has bought upon a new age is further solidified by the music production post internet streaming.

The essay [7] is about how music in internet age has given rise to new genres as well introduction of new type of music. The deep penetration of internet in isolated societies has enabled them to get their voice which has resulted in generation of new options for music community to fathom in. This in turn has made music industry a very diverse pool of talent and taste with very high genre.

In [8] “Predicting Movie Success Based on IMDb Data” the research paper describes on how to increase the revenue from a particular movie based on whether a movie is likely to succeed or fail. It employs a variety of supervised learning methods to determine if the cost of advertising can be reduced if the prediction is the movie will fail. Analysis of news segments and social media networks is suggested as a potential scope. The model is mainly for functional purposes and cannot be deployed in real-time, which is a drawback of this study.

In [9] ”IMDb Explorer: Visual Exploration of a Movie Database”,the IMDb Explorer, a web-based visualization tool that consists of two main views denoted by the movie universe and the career lines. The paper focuses on detecting associations between movies and the actors in those movies. Also, the analysis of time-varying data patterns like trends, counter-trends, anomalies, and outliers is a main component of the paper. They plan to expand the tool in the future by incorporating more interaction techniques and attempting to explore new application situations, such as paper titles contained in the DBLP.

In [10] the author view reading is still the most common leisure activity for the most of the people, and it continues to provides a unique path of the knowledge and learning. As a result books continue to be an important cultural material that is widely consumed. Despite the fact that over 3million are released every year, only small percentage of them are widely read, with fewer than 500 making it to the New York Bestsellers lists. Just a few authors are able to command once they arrive.

[11] The existence of the New York Times, and the inside story of its effort to navigate the new world, embracing the intensity of the web while remaining committed to reliable reporting and analysis, is at the root of the problem.

III. Methodology

A. Data Acquisition

Dataset 1. TMDb movies data

Source: https://api.themoviedb.org/3/discover/movie

There are various APIs available for movies data for example, IMDB, OMDB and TMDb. Out of these TMDB is freely available and easy access data from. The Movie Database (TMDb) API is a movie and television database created by the public since 2008. TMDb has been a premier source for metadata, with over 400,000 developers and businesses using the website. TMDb is a database that is used in over 180 countries. The API has different collection of movie titles with respect to genre, cast, popular and so on. The data-set uses “Discover” API from TMDb. The data-set selected for the purpose of this project has movies by various forms of information such as average ranking, date of release, number of ballots, genres, and certifications. Here using an API key 500 pages which gave an output of 9980 records and 11 columns were selected.

Dataset 2. Spotify Music data-set

Source:https://developer.spotify.com/documentation/web-api/reference

The Spotify Music API is one of a user-friendly API which has a vast collection of different playlists. It adds songs which are trending to the playlists. As this was a basic requirement for the music data-set, Spotify Music API was chosen.Spotify is a Swedish music listening and video services company based in Stockholm. Spotify is used in much of Europe and the Americas, as well as Oceania, Africa, and Asia, including South Africa and Mauritius. The API has a catalog of millions of songs sorted by mood (Dance-ability, Valence, Energy, Tempo), properties (Loudness, Speechiness, Instrumentalness), context (Liveliness, Acousticness), segments, beats, and other audio features. Here 15242 records from the Spotify API which has different fields such genre, title, year, album name, artist name, artist id etc. were selected.

Data-set 3. Multi-platform Video-games Database

Source: https://api.rawg.io/api/games

Considering the data-set for videogames, RAWG API was chosen as it has more than 350,000 games for 50 platforms including mobile platform. An added advantage to choose this API was that there was ample information regarding the genre, publishers, release dates, player activity data and Metacritic ratings. The data-set contains a vast selection of video-games from various formats and genres from the year 1971 to 2018. The original data is sourced through an API in JSON format. RAWG is a robust online database that allows users to manage and organize video-game records. The data retrieved from the API includes fields including title, genre, year of release, system requirements, and critic score. 1000 pages were sampled to retrieve data with 20000 title entries for this research paper's intent and study.

Dataset 4. NYT Bestsellers Database

Source: https://www.kaggle.com/cmenca/new-york-times-hardcover-fiction-best-sellers

For books there was a huge struggle to find an API which would give more than 5000 records along with its details such as title, genre, ratings, readers reviews, authors details and other details. So a JSON file from Kaggle was considered. The data-set is originally acquired from the New York Times API in Kaggle. The JSON file format is considered for this project. The book title, author, date of the bestselling list, published date of the list, book summary, rating (this week and last week), publisher, number of weeks on the list, and price are among the information gathered. It has 10000 records along with the other details.

B. Data Management

The collected data would be unstructured and in JSON format. There would be a problem with saving the records of all the data-sets. Mongo Db is the easiest way to manage unstructured data. Mongo Db is a document-oriented NoSQL database. It can be used to store both structured and unstructured data. The initial step is to establish a connection, creating collection and then inserting data. Mongo Db was used platform to store the data collected from different API sources for all four categories of movies, music, books, and video-games. Once the data is stored it is retrieved, analysed and pre-processed using Python. The clean data is then stored in PostgreSQL for further analysis and visualizations after merging all the four data-sets.

C. System Flow Diagram

Figure 3 shows a machine flow diagram that visually depicts the process flow in the project.

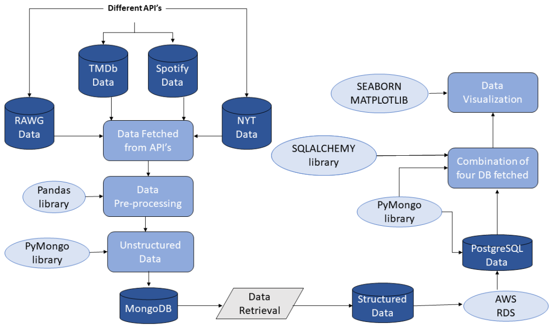

Fig. 3. System Flowchart

The project's ultimate aim is to examine and highlight various forms of entertainment that the general population enjoys. To do this, multiple data-sets were collected from APIs and downloaded from websites, each focusing solely on entertainment sources. These semi-structured databases, in which the data is unstructured, would be stored in MongoDB. To process the unstructured data, it will be fetched back in python to perform cleaning operations. The final data frames will be stored on a remote PostgresSQL server where appropriate visualizations will be showcased to match adjacent elucidations of the whole premise of project.

D. Data Pre-processing

1) Data Extraction:As previously mentioned, the four data-sets considered in this project were derived from various API sources. Using the "pymongo" library, data is retrieved from APIs and saved in JSON format. PyMongo is a Python library that includes methods for communicating with the MongoDB database. After the installation and importing of required sets of libraries, connection is made with MongoDB. Parsing through several pages to retrieve data since the content is spread over several pages and 5000 records were needed. Iteration for a loop to fetch data greater than or equal to 5000 records, considering the number of records per page. For example in case of video-games data-set nested absent values, try:except loop was implemented to find out non existing entries in the data. They were converted into 0 and consequently removed. These data-sets are then transformed into data frames with pandas library. Python is used as a programming language as it has a large library of open source data processing resources that are both usable and readable.
2) Data Cleaning:It means detecting and fixing corrupt or inaccurate information from a record collection, table, or database by finding missing, wrong, inaccurate, or meaningless portions of the data and either adding, updating, or removing the dirty data. In these data-sets,

fields were converted from one data type to another where appropriate, examined null values in various fields, excluded duplicates, and eliminated irrelevant data. Each data-set's date column was in DD-MM-YYYY format which has been converted to a required format, as $release\_date$ has been converted into date time format using $pd.to\_datetime$ function. ”Year” was removed from the conversion, as it was needed to look at the pattern in each group with respect to year. Columns were removed having zero values from the newly derived “Year” column.

3) Data Transformation:The method of transferring data or information from one format to another, typically from the format of a source system to the format required by a new destination system, is known as data transformation. The new column “Year” in each of the four data-sets will be a serving a primary motive in the analysis of trend for each category of entertainment namely movies, music, video-games, and books

E. Data Storage

Data was stored in two phases:

1) **Storing unstructured data**: The data fetched from the respective API's have been stored in JSON format into MongoDB. MongoDB is a document database with a horizontal scale-out design. Each row in a MongoDB database is a document defined in JSON, a formatting language, rather than tables with rows or columns as in SQL databases. MongoDB is an open-source software. MongoDB is an excellent way to store data because of many of these features. Inserted fields are supported by JSON documents, allowing relevant information and data arrangements to be stored with the document rather than in an outer table. Document databases are highly adaptable, allowing for changes in document layout as well as the handling of partially completed records.

   The pymongo library is used to create the mongo DB relation. Exception handling is considered when connecting to a database. Following the establishment of the connection, the JSON format data of movies, music, video-games, and books respectively are stored in separate collections objects. It is checked if the database exists prior to the insertion of data into MongoDB. If the database already exists, it will be dropped, and new insertion of the data will take place in the collections namely movies, music, test1 and books.

   The data extraction stage will then introduce data-sets into the recently created data frames. Every row is uniquely defined by a field named id. The data from the data frames is cleaned and transformed as per the requirement for all the four data-sets. MongoDB generates a column called id by default which is not required for this projects scope. So, the id column is dropped along with few not required columns in the datasets after pre-processing. At this stage, the unstructured data in JSON has been converted into structured data which is fed to PostgreSQL.

2) **Storing structured data**:PostgreSQL is an open-source object-relational database framework with a lot of capabilities [3]. Firstly, a Postgres server connection needs to be established to store data in PostgreSQL. This is done using python code with the help of Psycop2 library. Psycop2 library tries to bind to the database after importing the adapter. If the connection fails, a print statement to STDOUT will be sent. Following a successful link, a new database called db_DAP is created in Postgresql using Python using Inner Join and Left Join for books dataset. After that, a database link is established in order to read all four data frames, having four different Postgres names named $Movies\_Data$ (TMDb data), $Music\_Data$ (Spotify data), $Games\_Data$ (RAWG data), and data(NYT bestsellers data). It makes use of the SQLAlchemy library, which creates tables and data-types automatically based on the data loaded into Postgres. Data from four different Postgres tables were combined and populated in the Panda Data Frame using the SQL UNION query. After this, visualizations are performed.

F. Data Visualization

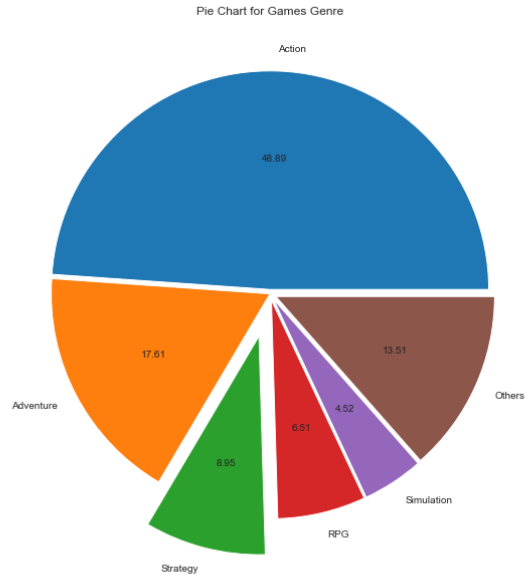

Fig. 4. Top 5 Games Genre

For performing data visualization and elucidating comparative information, python functions from Seaborn and Matplotlib libraries were imported. Seaborn is primarily preferred for statistical visualization due to its inbuilt ability to project stunning visually appealing pictorials. Attractive visuals are soft on the eyes and make audience take interest. In this report, seaborn library was called to implement heatmap to discuss the correlation between different parameters of the combined datasets. Another implementation was used for the construction of line plot chart adjacent to years. A comparative graph to compare growth is also plotted.

Figure 4 depicts the top 5 genre contribution to overall game released.

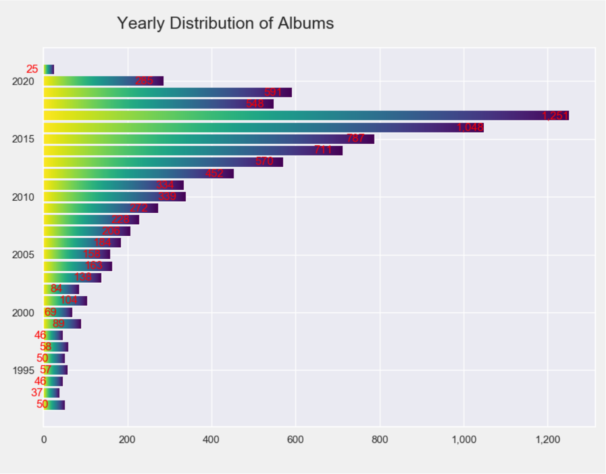

Fig. 5. Album Yearly Distribution

Figure 5 plots the yearly distribution of music albums released.

Matplotlib is a python plotting library which is designed to resemble MATLAB. Although the resemblance is in interface only and it is use as an alternate is discouraged. In this report matplotlib was called to construct a bar graph for the combined datasets. For genre, a violin plot was constructed for each dataset across years to see any data is worth elucidating regarding the popular choice. A comparative analysis is predicted to give insight among the different consumer groups.

In figure 6 the bar graph showcases the yearly distribution of movies released.

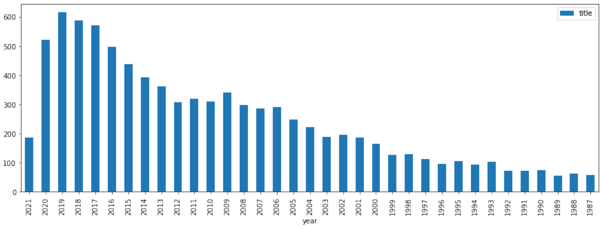

Fig. 6. Yearly Release of Movies

IV. Results

The heatmap in Figure 7 showcases the correlation between the year and entertainment source categories. Color coding ranges from black to orange- hued white. It is clearly visible that books have negative correlation with the period. As predicted the book industry has suffered with the passing years. It is in stark contrast with other source of entertainment which share highly positive correlation with period. A special mention to Movies and Gaming for unexpectedly high correlation with year which concludes towards the direction that Internet has affected Movies, Music and Gaming for the better.

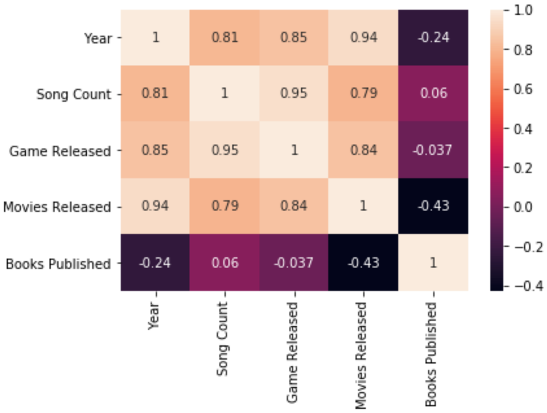

Fig. 7. Correlation Matrix

Figures 8 and 9 showcase the yearly count of all 4 entertainment sources with respect to each other. These are important visualization to ponder upon as it reflects the true motive of the project. On individual inspection yearly growth of all datasets differ in magnitude and trend. Videogaming shows significant growth with unexpectedly high growth numbers from 2013 on-wards. This is in line with growth of internet and onset on online gaming as well the positive achievements in computer technology with much powerful devices being available to general populace. Software developing companies have also invested hugely in videogaming divisions which has resulted in focused developing centers publishing video games at a higher rate.

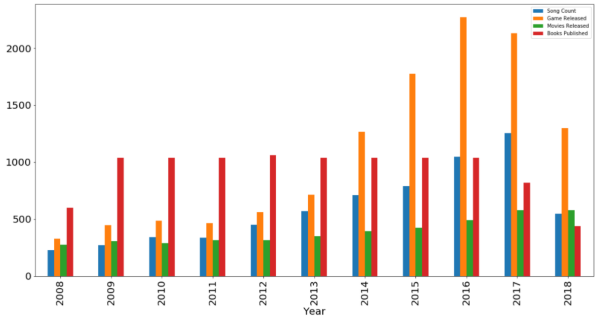

Fig. 8. Multiple Variable Bar Graph

In contrast the Books published during the same period show no growth at all. In fact, a slight decline can be seen at the concluding years. The print media industry is seen facing challenges as more populaces move towards a digital platform. An important information to keep in mind while visualizing this plot is the absence of e-books data used.

The growth of Music and Movies Industry is stable. This shows the evolving nature with which the stakeholders in these

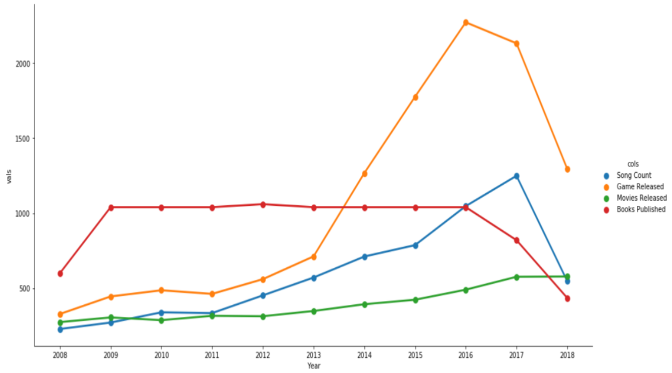

Fig. 9. Line Chart

sectors have undertaken with regards to penetration of Internet among the populace.

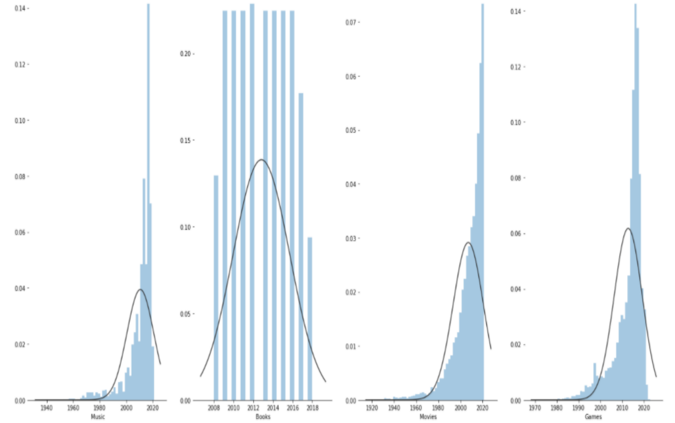

Fig. 10. Distribution graph

In figure 10, the data represents the yearly distribution of count for different scale of y-axis. Special emphasis should be placed on each individual chart in their respective x-axis. The yearly distribution of releases is in-line with last three visualizations.

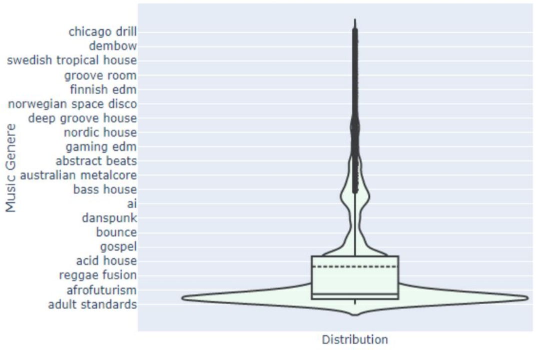

Fig. 11. Violin Music Genre Plot

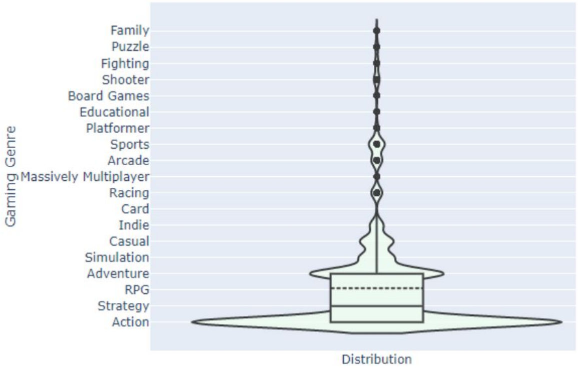

Fig. 12. Violin Games Genre Plot

In figure 11,12,13,14 the visualization contains violin plots of all 4 entertainment sources. The graph represents the distribution of content across genre among all the data present in the set. This is essential to study the popular choice of genre that the consumers enjoy. Music and video-games consists of many genres but the market is captured by only few. The Action and Strategy genre enjoy greater popularity while there is hardly any presence from other genres. The same can be said about the Music industry where adult standards and Afro-futurism have lead.

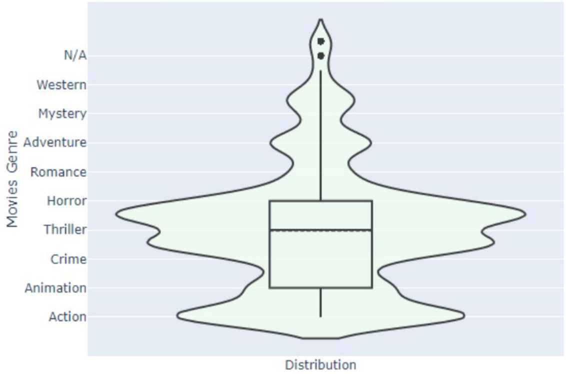

Fig. 13. Violin Movies Genre Plot

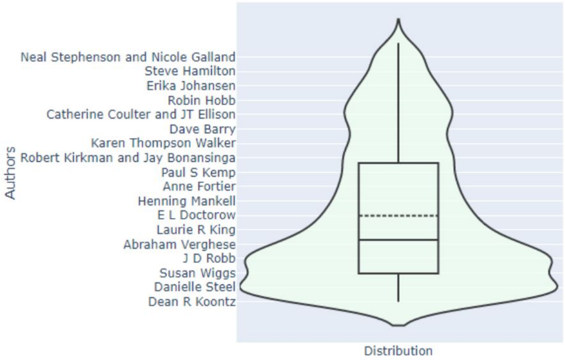

Fig. 14. Violin Books Author Plot

Movies and Books have fairly evenly distributed releases among genres. The Thriller and Horror genre is still preferred over other genres but production in genres cannot be said to be lagging behind.

V. Conclusion

The report shares limelight on the impact the Internet has on the entertainment the general public enjoys. Industries which have digitized their marketing and sales pitch have benefited from this phenomenon. For any serious business entity to remain relevant in the current market, the knowledge of internet know-how has become necessary. This compulsion has forced even traditionally non-digitized platforms such as movies and music to evolve their ways.

Following the research question the report has successfully showcased trends in each entertainment domain. As expected, non-digitized forms of entertainment suffer from loose sales and profit. Highly interactive fast entertainment sources like video-games especially online games have enjoyed greater success. For print media industry like books and magazine, paperback prints have had stale market distribution.

The report has also shed light on the inner choice trends in each specific entertainment domain. While games and music have formed stronghold genres, movies and books have evenly distributed consumer base. Which trend is more beneficial is open to further study in future.

While the report has been successful in predicting isolated objectives, further research needs to be done on additional sources of entertainment and their market share. A new generation of content in the form Netflix, Amazon Prime, Voot and hotstar shows and movies have propped up and garnered a fair share of slice in the entertainment market. The study also does not take any consideration of the live performance and shows that need personnel attention on behalf of consumers.

The report showcases how internet has changed the playing field for entertainment sector. This can be gauged in organized systems and works need to be done to make existing platforms digitized to remain relevant. The Internet has brought in a new age and stake holders can either evolve and reap the benefits or be left behind in oblivion.

References

[1] RAWG Video Games Database API , Accessed:April 26,2021.[Online]. Available:https://api.rawg.io/api/games.

[2] Nuremberg Institute for Market Decisions, ”What are our sources of entertainment?”, Accessed: April 1, 2021.[Online].Available: https://www.nim.org/en/compact/focustopics/what-are-our-sources-entertainment.

[3] Tutorialspoint,”PostgreSQL Tutorial”, Accessed: March 30, 2021.[Online]. Available:https://www.tutorialspoint.com/postgresql/postgresql.

[4] Torben Grodal,”Video games and the pleasures of control,” published by University of Copenhagen,publication 79, dated October 2012.

[5] Peter Zackariasson, Timothy L. Wilson, “ The Video Game Industry:Formation, Present State, and Future,” published by Routledge, 2012.

[6] Jim Rogers, ”The Death and Life of the Music Industry in the Digital Age,” published by A& amp;C Black, 9 May 2013 - Social Science.

[7] Steve Jones, ”MUSIC THAT MOVES: POPULAR MUSIC, DISTRIBUTION AND NETWORK TECHNOLOGIES”, Accessed: April 1, 2021.[Online].Available: https://www.tandfonline.com/doi/pdf/10.1080/09502380110107562.

[8] Nithin VR,, Pranav M, Sarath Babu PB, Lijiya A,”Predicting Movie Success Based on IMDb Data,” published by National Institute of Technology Calicut, July 2014.

[9] Michael Burch, ”IMDb Explorer: Visual Exploration of a Movie Database,”volume VINCI’18, August 2018, Växjö, Sweden.

[10] Burcu Yucesoy, Xindi Wang, unming Huang,Albert-Laszlo Barabasi, ”Success in books: a big data approach to bestsellers,”Volume ecember 2018EPJ Data Science 7(1).

[11] Jeff Gomez, ”Print Is Dead: Books in Our Digital Age”.