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 Videogames 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. Videogaming has seen tremendous growth and in turn benefitted the most from online campaign by developers. Print media has a strong foothold and is predicted to shift toward digitization in the near future.

Keywords – Entertainment, music, movies, books, videogames, 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 we have 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,we have chosen Music, Movies, Video games and Books as the study areas for our 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 eBooks. 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 videogames 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.

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. 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.

The research question that needs special mention is:

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

II. RELATED WORK

In [1], 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 videogames showcases that binds the user and hence establish itself as viable source of entertainment.

[2] is a full comprehensive collection of all the information about videogame. The book discusses in depth about the

growth of videogaming 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 [3] 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 [4] 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.

A. Maintaining the Integrity of the Specifications

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 below figure 1 shows the data extracted from the API.

Fig. 1. JSON format Movies data

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 our 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 dataset uses "Discover" API from TMDb. The dataset 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 we have selected 500 pages which gave us an output of 9980 records and 11 columns.

Dataset 2. Spotify Music dataset

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 dataset, we chose the Spotify Music API. The

Fig. 2. Movies data columns

data retrieved from the API is as shown in the figure 3.

Fig. 3. JSON format Music data

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 (Danceability, Valence, Energy, Tempo), properties (Loudness, Speechiness, Instrumentalness), context (Liveness, Acousticness), segments, beats, and other audio features. Here we have selected 15242 records from the Spotify API which has different fields such genre, title, year, album name, artist name, artist id etc. as seen in figure 4.

Dataset 3. Multiplatform Videogames Database

Source: https://api.rawg.io/api/games
Considering the dataset of videogames as shown in below figure 3, we have chosen the RAWG API as it has more than 350,000 games for 50 platforms including mobile platform. An added advantage to choose this API was that we had information regarding the genre, publishers, release dates, player activity data and Metacritic ratings.

The dataset contains a vast selection of videogames from various formats and genres from the year 1971 to 2018. The

Fig. 4. Music data columns

Fig. 5. JSON format Videogames data

original data is sourced through an API in JSON format. RAWG is a robust online database that allows users to manage and organize videogame records. The data retrieved from the API includes fields including title, genre, year of release, system requirements, and critic score, as seen in figure 6. We sampled 1000 pages of data with 2000 title entries for this research paper's intent and study.

Fig. 6. Videogames data columns

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, we have considered a JSON file from Kaggle. The data obtained from the file is depicted in the figure 7.

The dataset is originally acquired from the New York Times

Fig. 7. JSON format Books data

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 information as shown in figure 8.

Fig. 8. Videogames data columns

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 datasets. 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. We have used Mongo Db platform to store the data collected from different API sources for all four categories of movies, music, books, and videogames. 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 datasets.

C. System Flow Diagram

Figure 9 shows a machine flow diagram that visually depicts the process flow in our project.
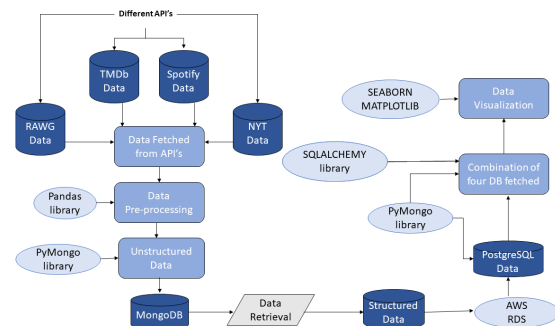

Fig. 9. System Flowchart

The project's ultimate aim is to examine and highlight various forms of entertainment that the general population enjoys. To do this, multiple datasets 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 datasets 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. We had to parse through several pages to retrieve data since the content is spread over several pages and we need more than 5000 records. We iterated the loop to fetch data greater than or equal to 5000 records, considering the number of records per page. For example in case of videogames dataset 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 datasets 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 datasets, we converted fields from one data type to another where appropriate, examined null values in various fields, excluded duplicates, and eliminated irrelevant data. Each dataset'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. We have extracted "Year" from the conversion, as we need to look at the pattern in each group with respect to year. We have removed records 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 datasets will be a serving a primary motive in the analysis of trend for each category of entertainment namely movies, music, videogames, and books

E. Data Storage

We have data 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, videogames, and books respectively are stored in separate collections objects. We check 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 datasets 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 datasets. 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. 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 datatypes 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.

IV. Visualizations

V. Conclusion

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [**?**]. Refer simply to the reference number, as in [**?**]—do not use "Ref. [**?**]" or "reference [**?**]" except at the beginning of a sentence: "Reference [**?**] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [**?**]. Papers that have been accepted for publication should be cited as "in press" [**?**]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].

References

[1] M. Shell. (2007) IEEEtran homepage. [Online]. Available: http://www.michaelshell.org/tex/ieeetran/

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.