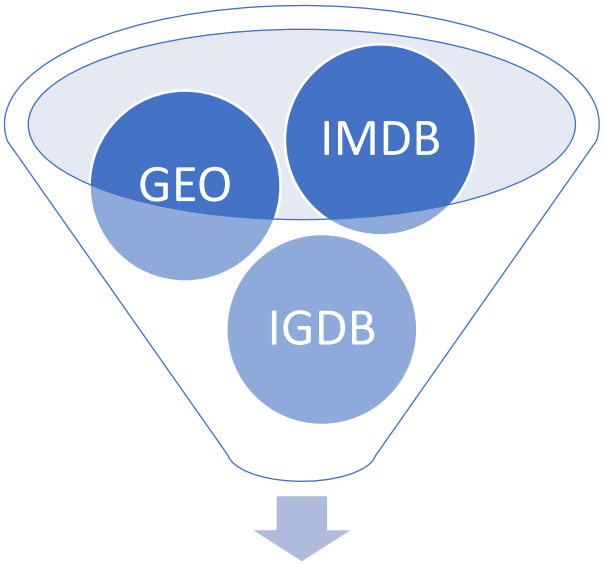
Video Game Publisher Information

Data-420 Project Report



Applied Data Science
University of Canterbury,
New Zealand.

Jian ZHOU - 51404140
Jiaying Zhu – 44820888
Shi Chen - 54638177
Waqas Naveed - 88354613
Xiaohong Chen - 24908339

21st October 2020.

Contents

Preface

The idea is to provide a platform to the game publisher company to find the relevant characters & the countries to launch the new game with their unique characters and storyline. The storyline is extracted from the imdb database and

We have gathered the data from multiple sources, wrangle it, join it and made a meaningful information which can then be used for game publication.

Tool & Data Source

- R
- Julia

We have extensively used both R & Julia to scrape, fetch the data APIs, wrangle and visualize the data. All the NOTEBOOKS ARE IN THE Source directory along with the downloaded/extracted data. The notebooks do have the information about the installed packages.

Data Sources

IGDB API

Scource link (IGDB API documents): https://api-docs.igdb.com/#endpoints
IGDB introduction: https://www.igdb.com/about

Why IGDB

- From the view of authority: the community and developers in this flat are generous, they wish to share data.
- Professional: IGDB is games ”one-stop-infospot”, which allow users to do more with the information in order to explore and find games to play in a new way. From the authority point of view. And it contains 245,500 games, more than 35,647 members and involved 27,894 companies. It is very professional API resources.
- It is free!

IMDB website

Source link = https://www.imdb.com

Why IMDB

- IMDb is the world's most popular and authoritative source for movie, TV and celebrity content to find ratings and reviews for the newest movie and TV shows.
-
- Scrap feature: IMDB does not provide the freeware API for automated queries and the free part is very limited. We are exploiting the powerful Julia code to scrap the site and get movie and actor characters from www.imdb.com.

CoinMarketCap

Source link: https://coinmarketcap.com/all/views/all/

Why CoinMarketCap

- The website 'CoinmarketCap' is famous in the cryptocurrency field, and it is well-structured
- The information is well structured and complete to fulfil out purpose.

Geo Data

Source link: https://stefangabos.github.io/world_countries/

Why GeoData

- looking for a suitable geo data source to combine with the existing one subset(game company) of IGDB to draw a map.
- It is an open source.

Targeted features

We have a distributed set of data which we are going to merge in order make a useful information. The wrangling process covers different features which are targeting as part of each dataset. The following features are categorized accordingly:

IGDB :

- Genre (market preference)
- Age_rating (target user)
- Developer map (parterners or competitors)
- Keyword pattern

IMDB:

- Movie
- Storyline
- Characters
- Locations

CryptoPayment:

- Coin’s name
- symbol
- Market cap
- Price
- Circulating supply

GeoData:

- Visualization of countries
- Location based data mapping

Challenges

Every part of data collection/wrangling has its own challenges which we came across on our journey of data wrangling. We have listed them according to their data sources and the tools we have used to extract the information.

IGDB

Token: To new hands, challenges are everywhere happening every day. Scaping data from API, the first thing is to obtain "Token" from API Database. It is the key. We Follow the process list on API document, doing it step by step, though totally have no idea what will happen next. With Gulio's help, we eventually can post require to IGDB API with Token.

API documents: There are a lot of information in the API documents, "authentication" "Example" "Endpoints" contains a lot of information's about the game dataset. We did not read them carefully in the beginning and made problem when doing the data cleaning. When we realized it, we already waste a lot of time. This is a big lesson. Whenever collection data, the first thing is figure out what kind of data need to be collected, it must be clear and reasonable.

Coding**:** We are using Julia to do this part. Not easy at all. From "post" to generate string information from web, then transform them into JSON format then output as CSV. The process seems very clear, but when you are doing it, there are a lot of small tricks waiting to stuck you.

Logic**:** This is a very important part. As mentioned before, in the very beginning, we generate some data, then we think it will work. But actually we use the wrong logic, the data cannot be used, too many notice data, missing value data. After we define what is "popular game", how to filter them directly rather than scaping all the games' data. Things become clear.

Correlation **:** Here, the correlation means the relationship between main dataset and sub dataset. In the main game information dataset, there are plenty encoding values cannot be used directly. This encoding values need another sub dataset to decoding. That is why we generate 10 sub dataset. And there are some features' type are list, it is another challenge part. We have not solve yet.

Github : Self-learn how to use Github. It is a challenge for us in the beginning, but it is quite convenient and smart after we familiar with it.

IMDB

Coding : We don't have an API to use for the movies and characters, and the IMDB has a very large dataset. For movies, we name, year, rating, votes, directors and cast member from two different web pages. The Programming languages used are R (for twitter API, sf map, NLP wordcloud2), Julia (for IMDB web scraping) and Java (for Clavin storyline locations extraction).

Big Data : There are millions of records for cast members, in addition to decades' movie and stars numerical and text data. IMDB contains around 10,000 movies for each year's record.

Techniques : NLP analysis to extract geographic locations from movie stories, and created wordcloud for characters and storylines from both IGDB and IMDB datasets.

Incrementally save data to files. :

Two steps approach:

1. Movie data with IMDb ID.
2. Scrape cast page based on movie table.

CryptoCurrency

The website 'CoinMarketCap' has API for python but not for R. I wrote two functions to get historical cryptocurrency data and merge the dataframe together into one dataframe.

To represent cryptocurrency trends in the last 12 months by days more clearly, We explored the gganimate to plot the data.

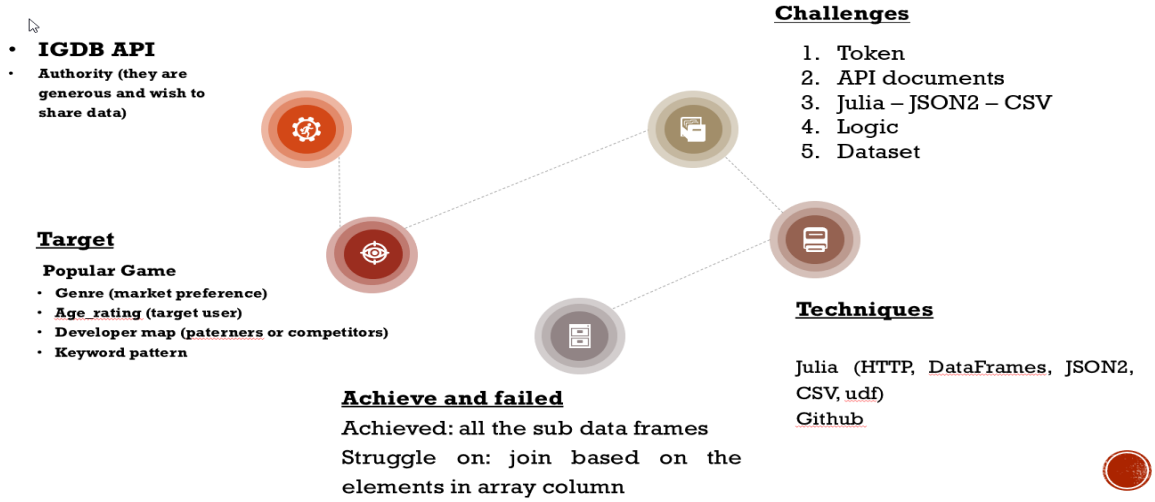
It is a little bit difficult for me to learn how to use github in the beginning. After some trying and our team members' help, I figure out how to use this platform.

GeoData

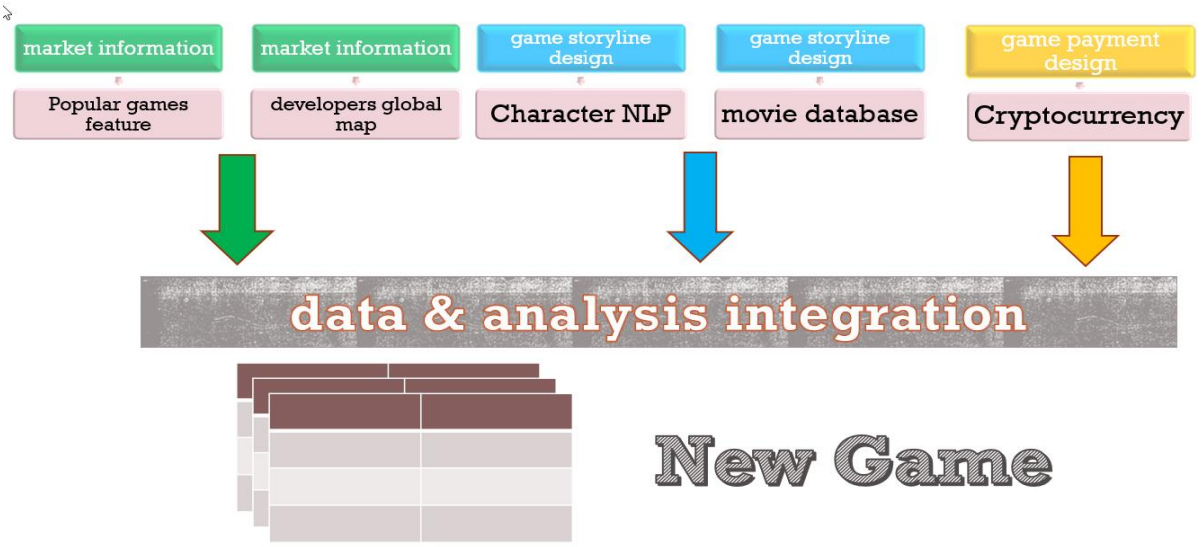
The company's world mapping is using the wrong data, but at least try to plot. By using R(ploty) it came to an issue for presentation in Jupyter lab.

Since we changed topic in the middle part of project, other geo resources could not be transformed.

Geo data API contains many limitations (such as Google Geo API, NZ Post API), so they are not be used in final project.



Dataflow Diagram



Achieve/Failed

We have collected massive amount of data from different sources and achieved a lot but there are few failures too. They all are listed categorically as per the data sources:

IGDB

Achieved

Our main target dataset (the main popular games data) was generating successfully (filename: target_games_fellow.csv). It is in the Data422-project/IGDB/Docs directory. This file contains the top 500 popular games and 25 game's features information.

refer to Data422-project/IGDB/Codes/IGDB_games_infor_scrap_main.ipnb section 1 keyword "body_fellow")

Another 10 sub datasets (decoding dataset related to these 500 popular games), which can be used to do different group analysis. Such as game_ageRating_groupCount_plot.png and game_ageRating_groupCount.csv is using the data from game_ageRating sub dataset. You can do more as you like, such as genre groupcount, releaseDates analysis, game_theme groupcount and so on.

Failed

1. Companies information sub dataset was generate in wrong coding. Have tried many ways but still not been solved yet.
2. The company's world mapping is using the wrong data, so please use it carefully.
3. Have not figure out Julia technique:
4. Convert list column into a set.
5. Join by list column.
6. Groupby id summarize column value as set.

IMDB

Achieved and Issues

To avoid ethical and privacy problems, only text mining fictitious characters instead of real names.

Finally gathered 0.3 million records cast member dataset, in addition to 10 years' movie and stars numerical and text data, around 15MB movie and stars data, as well as a 12MB cast table, not including the text files for storyline and characters, which are also big enough to do text mining. This cast table especially can be used for other data scientists for other purposes, such as to find the most valuable actors / actresses throughout the years, when combined with movie and / or stars table, which possess rating, score, voting etc.

NLP analysis to extract geographic locations from movie stories, and created wordcloud for characters and storylines from both IGDB and IMDB datasets. CLAVIN on GitHub to extract geographic information from text files, but went through a long process to install it on local computer.

FlipTextAnalysis library in R to do NLP sentiment analysis in wordcloud2.

rtweet API to get game and movie text and wrote code snippets to filter out stop words.

Simple Features (sf) R library combined with a world geojson file I found online to plot storyline locations in 2010 and 2020 and found that India, China, France, Mexico and Japan are some of the hotspots. The popularity of India continues but China decreases.

Open Source projects are sometimes not so perfect in case of Clavin producing some unnecessary data, making the output untidy, but it's a nice attempt. The graph output still gives us some intuitive peeking into the data we managed to retrieve from both IGDB and IMDB.

CryptoCurrency

Achieved

1. Top 200 cryptocurrency information including coins' name, symbol, market cap, price, circulating supply, volumn 24h, change 1h, change 24h and change 7days.
2. Found 'Tether', 'Bitcoin', and 'Ethereum' are very popular recently by using ggplot.
3. Obtained top5 cryptocurrencies data in the last 12 months and merge the dataframe together.
4. Using gif plots to illustrate the data and find patterns.

Failed

1. Web scraping 2000 cryptocurrency data at once from 'CoinMarketCap'.

GeoData

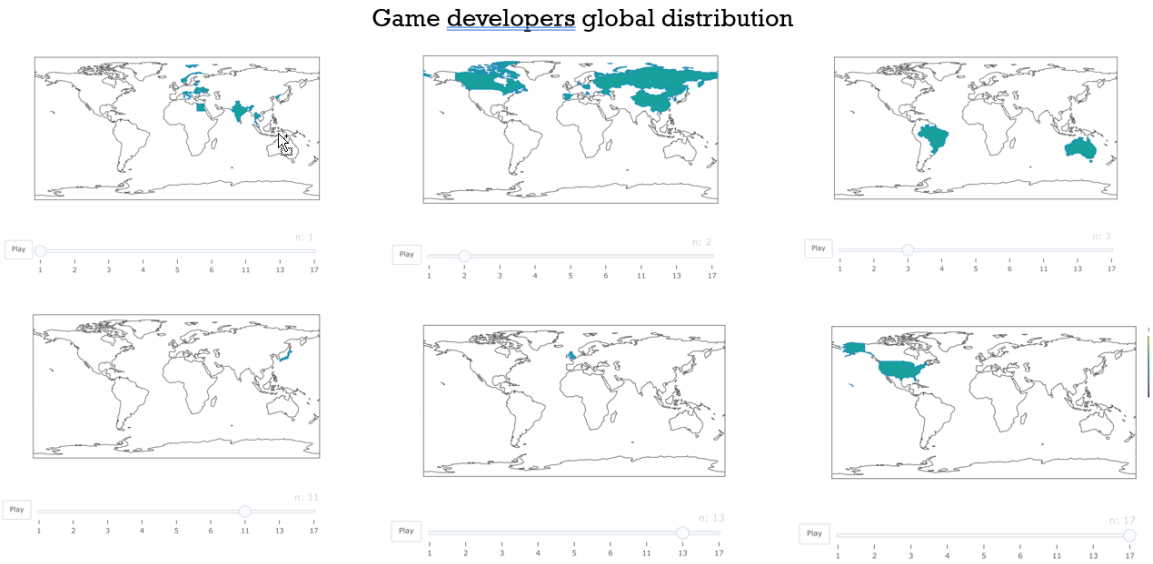
Achieved

Geo selection please refer to Data422-project/Geo/Codes/geoData.ipynb (2.connect datasets for visualization), there is one global map to view the distribution of 21 countries and one bar chart to view the company(developer) counts in each countries.

Failed

1. The companies world mapping is using the wrong data, but at least try to plot. By using R(ploty) it came to an issue for presentation in Jupyter lab.
2. Since we changed topic in the middle part of project, other geo resources could not be transformed.
3. Geo data API contains many limitations (such as Google Geo API, NZ Post API), so they are not be used in fianl project.

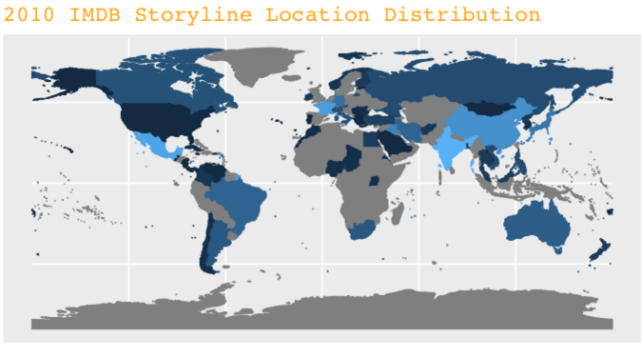
Data Visualization

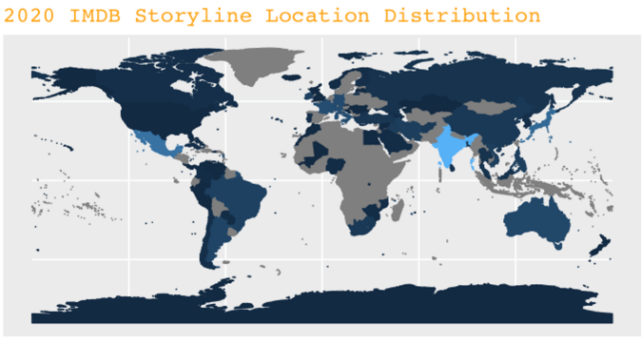


1, Data Dimensions Selection →

2, Wrangling : Games → Companies (Developers) → Countries → Geo Data →

3, Visualization : Plotly's R : *plot_geo(df)*





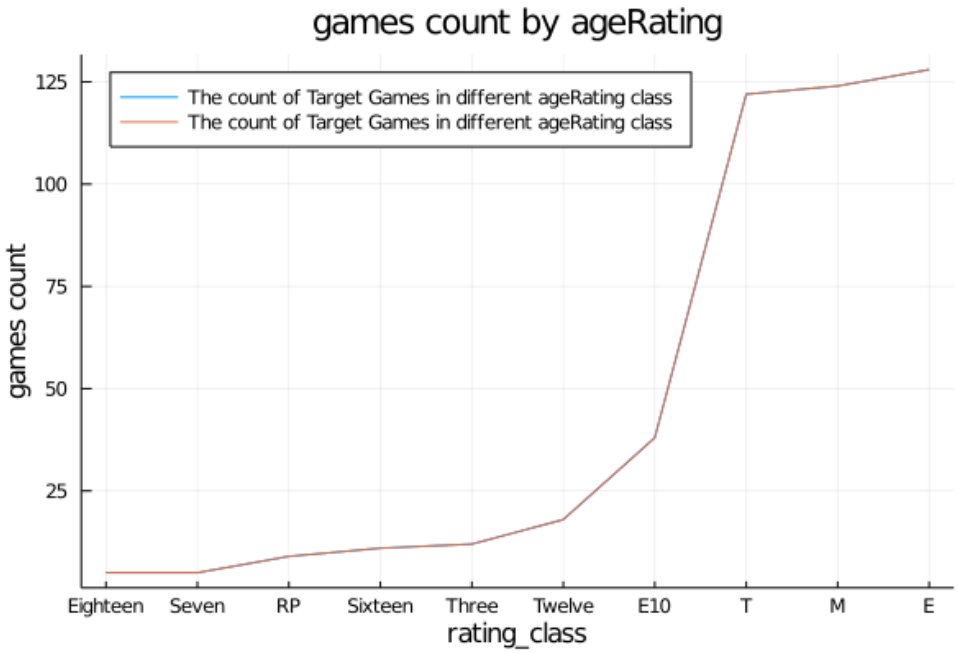
IMDB Storyline Location Distribution

2010

VS

2020

Games Storline keywords:



References:

The notebook contains code are placed under GIT repository along with other material including images and data downloaded either from API or scraped using web. :

~\ Data422-project\IGDB

~\ Data422-project\IMDB

~\ Data422-project\Geo

~\Data422-project\Cryptocurrency

Project report along with other textual information is placed under

~\Data422-project\Docs