A picture containing text, clipart

Description automatically generated

*School of Computing*

|  |  |
| --- | --- |
| Student Name | Sakthignana Sundaram Somaskandan |
| Student Number | 14346091 |
| Email Address | Sakthignana.somaskandan2@mail.dcu.ie |
| Program of Study | M.Sc. in Computing (Part-time) |
| Programme Code | MCM |
| Project Title | Data Visualisation Assignment |
| Module code | CA682I Data Management and Visualisation |
| Lecturer | Dr Suzanne Little |
| Project Due Date | 6th December 2022 |

|  |
| --- |
| *I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of other in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.*  *I/We have read and understood the referencing guidelines found at* [*http://www.dcu.ie/info/regulations/plagiarism.shtml*](http://www.dcu.ie/info/regulations/plagiarism.shtml)*,* [*https://www4.dcu.ie/students/az/plagiarism*](https://www4.dcu.ie/students/az/plagiarism) *and/or recommended in the assignment guidelines* |

Name: *Sakthignana Sundaram Somaskandan* Date: *6th December 2022*

Table of Contents

[Abstract 3](#_Toc120054282)

[Data Collection 3](#_Toc120054283)

[Data Exploration, Processing, Cleaning and/or Integration 4](#_Toc120054284)

[Game statistics 4](#_Toc120054285)

[Review sentiment 5](#_Toc120054286)

[Data exploration 5](#_Toc120054287)

[Visualisation 5](#_Toc120054288)

[Conclusion 5](#_Toc120054289)

[References 5](#_Toc120054290)

# Abstract

Max 200 words

What is the question you are answering or the story you are trying to tell?

What is the conclusion that you reached?

# Data Collection

½ page

Where/how did you retrieve your data? Provide a URL if available online.

Describe the data – size (GB or attributes), number of rows, attributes, data types present

What aspects of big data (volume, variety, velocity) are present in your dataset(s)?

I downloaded my dataset from [Kaggle](https://www.kaggle.com/) called [Steam Reviews Dataset 2021](https://www.kaggle.com/datasets/najzeko/steam-reviews-2021) which is in csv format. The dataset comprises of around 21 million user reviews (rows) of around 300 different games on Steam with size of 8.17 GB. Aside from the dataset obtained from Kaggle, I engineered and collected my own datasets using certain features from the Kaggle dataset, which will be discussed in detail under Data Exploration, Processing, Cleaning and/or Integration section.

The dataset contains the following attributes and data types.

|  |  |
| --- | --- |
| Attribute | Data type |
| Index | Number |
| Steam app ID | Number |
| App name | String |
| Review ID | Number |
| Language of review | String |
| Review text | String |
| Review creation timestamp | Number |
| Review latest update timestamp | Number |
| Whether review recommends the app | Boolean |
| Number of “helpful” votes for review | Number |
| Number of “funny” votes for review | Number |
| Score based on number of helpful votes | Number |
| Number of comments for review | Number |
| Whether review author purchased the app on steam | Boolean |
| Whether review author received the app for free | Boolean |
| Whether review was written during early access | Boolean |
| Review author steam ID | Number |
| Number of games review author owns | Number |
| Number of lifetime app reviews by author | Number |
| Author lifetime playtime of reviewed app | Number |
| Author playtime of reviewed app in last 2 weeks | Number |
| Author playtime of reviewed app at time of review | Number |
| Author time last played reviewed app | Number |

The three aspects of big data are present in my dataset in the following ways.

Firstly, it has variety: I used the app\_id column (Steam specific game identifier) from the Kaggle dataset to fetch game statistics on each of the 300 games as extra data to gain a deeper understanding.

Additionally, the game statistics data is updated monthly and made available through the same API endpoints that I’ve used. This fulfils the velocity aspect as the data is in motion.

Lastly, the Kaggle dataset contains 21,612,444 reviews and the dataset that I created contains 21,288 rows of monthly game statistics which is of size 1.8 MB.

# Data Exploration, Processing, Cleaning and/or Integration

½ pages

What did you need to do to prepare the dataset(s) to create your graph/chart?

How did you choose the attributes and data subset to visualise?

The story that I want to convey to the viewer required data that is not readily available in the dataset. Therefore, I had to compile my own dataset using certain information from the Kaggle dataset.

## Game statistics

The API endpoints I used were made available by Steam on RapidAPI, which is an online API marketplace. Developers use RapidAPI to discover and connect to thousands of APIs. I made calls to the following endpoint, with unique IDs fetched from the Kaggle dataset.

|  |
| --- |
| https://steamcharts.p.rapidapi.com/api/v1/games/{id} |

I was able to collect the following information from executing the above calls.

|  |  |
| --- | --- |
| Attribute | Description |
| ID | Steam app ID |
| Name | Game name |
| Month | Month and year associated with the statistics provided |
| Average players | Average number of players during a specific month |
| Gain | Number of players joined or left the game since the last month |
| Gain in percent | Number of players joined or left the game since the last in percent |
| Peak players | The greatest number of players in the game at the same time during a specific month |
| Playing 12-hours ago | Number of players in the game 12 hours ago |
| 24-hour peak | The greatest number of players in the game at the same time in the last 24 hours |
| All-time peak | The greatest number of players in the game at the same time over the game’s lifetime |

## Review sentiment

## Data exploration

# Visualisation

1-2 pages

Screenshot or image of visualisation.

Explain your choice of chart or graph type – what relationship or data type are you showing?

Design choices – justify your use of colour, shapes, marks, layout, structure, font, labels referring to books or articles as necessary.

Comment on any interactivity or animation and how it helps answer your question.

Give a list of tools or libraries used.

# Conclusion

½ - 1 page

Critically analyse the outcome of your visualisation with respect to your question or story.

Were there aspects that you think could be improved upon?

Were there effects or functionality that you were technically unable to achieve that would improve your visualisation?

# References

Include any citation of the dataset.

Include links to any tutorial or example that contributed significantly to your work.

Include any books, articles or web resources supporting your design choices.