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, 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

Table of Figures

Abstract

Product reviews have always been used to check the quality and legitimacy of a product, and product recommendations have played a vital role in creating big profitable corporate companies. I am answering whether the sentiment expressed in game reviews impacts player engagement. Despite having limited time and data, I can prove my point and illustrate the findings in visualisation to communicate the results effectively to the audience. I concluded that review sentiment (measured by transforming the textual review to sentiment scores) and player engagement (measured by average player count) are related.

Data Collection

I downloaded my dataset from Kaggle, called Steam Reviews Dataset 2021, which is in CSV format [1]. The dataset comprises roughly 21 million user reviews (rows) of around 300+ games on Steam with a size of 8.17 GB. Aside from the dataset obtained from Kaggle, I engineered and collected other datasets using specific features/columns 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 for which data is available as extra data points to gain a deeper understanding of player engagement.

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

Lastly, it has volume: the Kaggle dataset contains 21,612,444 reviews. The dataset I gathered contains 21,288 rows of game statistics of size 1.8 MB.

Data Exploration, Processing, Cleaning and/or Integration

The story I wanted to convey to the viewer required data that is not readily available in the dataset. Therefore, I had to collect other datasets using certain information from the Kaggle dataset, namely game ID. The Game statistics and Sentiment analysis sections below outline the approach and the data points obtained.

Game statistics

Steam published the API endpoint I used on RapidAPI, an online API marketplace. Developers use RapidAPI to discover and connect to thousands of APIs. I made calls to the following endpoint, with unique game/app 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 call.

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 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 in the last 24 hours
All-time peak	The greatest number of players in the game over the game's lifetime

The collected dataset needed to be cleaned for visualisation, i.e., the 'Month' column had 'Last 30 days' as entries, whereas the rest of the data are in the format [Month] [Year]. Therefore, I used pandas (python data analysis library) to replace all instances of 'Last 30 days' with 'October 2022'.

The Jupyter notebooks can be found in the GitHub repo as well as in the submission zip file.

Sentiment analysis

I decided to use the review text provided by the Kaggle dataset to compute the sentiment score associated with the text. I made use of the TextBlob library to carry out the sentiment analysis. TextBlob is a python library for processing textual data and provides a consistent API for common Natural Language Processing (NLP) tasks [2]. The sentiment analysis API returns a named tuple of the form `Sentiment (polarity, subjectivity)`. The polarity score is a float within the range `[-1.0, 1.0]` where `-1.0` is entirely negative and `1.0` is altogether positive. The subjectivity is a float within the range `[0.0, 1.0]` where `0.0` is very objective, and `1.0` is very subjective. I used the polarity score for visualisation as it shows the overall sentiment of a review.

Note: I performed sentiment analysis on English reviews only and filtered non-English reviews out of the dataset for sentiment analysis.

Sentiment Analysis Pseudocode

```
Read in the Kaggle dataset
Filter out unnecessary columns
Filter out reviews that are not in English using ‘language’ column value
Clean the review text by removing stop words and unnecessary characters
Perform sentiment analysis on the cleaned review text
Collect the polarity values
Append the polarity values to the Kaggle dataset dataframe
Export to csv file
```

I collected a polarity score in the range [-1.0, 1.0] for each review in the Kaggle dataset. This dataset contains the following information.

Attribute	**Description**
Steam app ID	Steam app ID
App name	Game name
Review ID	Identifier for the review text
Language of review	Language used in the review text
Review text cleaned	The cleaned review text – removed stop words, punctuations and other unnecessary characters
Polarity	The sentiment score in the range [-1.0, 1.0] indicating the review emotion, i.e., positive, negative, or neutral
Review creation timestamp	The date/time when the review was posted

The Jupyter notebooks can be found in the GitHub repo as well as in the submission zip file.

Data exploration

The charts below are not the final visualisations. I find it easier for me to get to know the data by visual cues; that is the reason why I included these graphs. I first explored the data by looking at the language ratio – the reviews written in English dominated the dataset, as seen in Figure 1. I also utilised a tree map to visualise the language split, as seen in Figure 2.

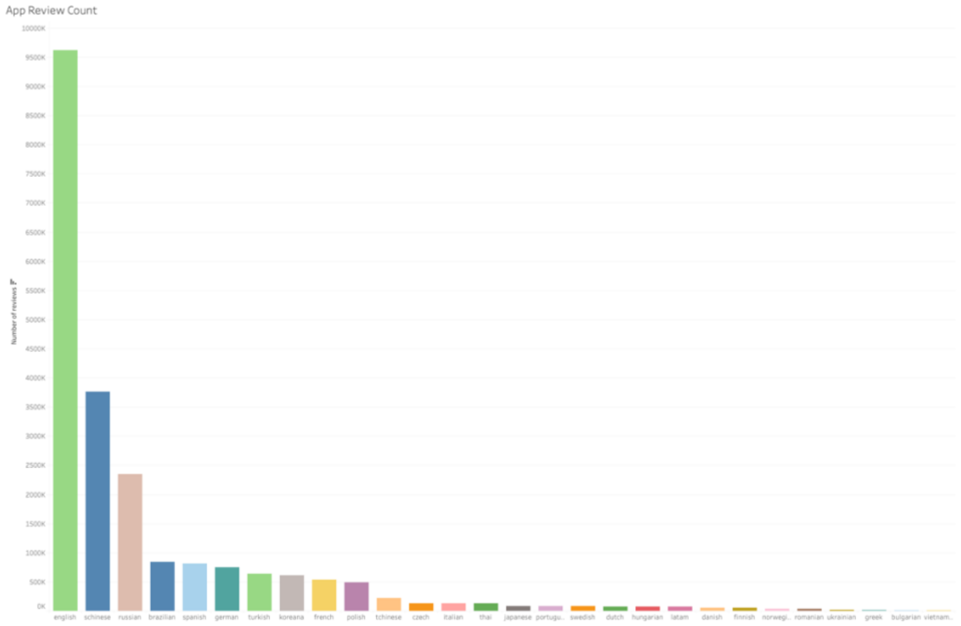

Figure 1: Language Ratio

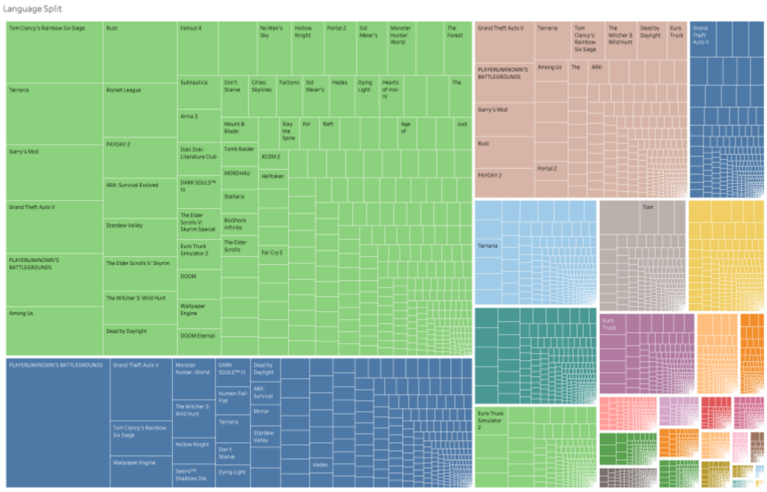

Figure 2: Language Split as a Tree map

This gave me the confidence that analysing the English reviews alone for sentiment analysis can capture a game's overall sentiment better than the other available languages.

Additionally, I explored the polarity split among the English reviews across all 300+ games in the dataset, as shown in Figure 3.

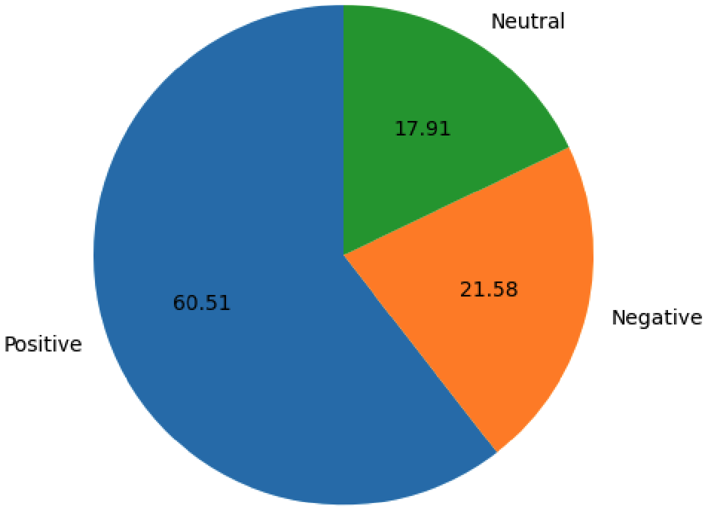

Figure 3: Polarity Split - Sentiment Analysis

Instead of strictly choosing attributes from one dataset, I gathered the data I needed, using specific attributes from the big dataset to convey my story effectively.

Visualisation

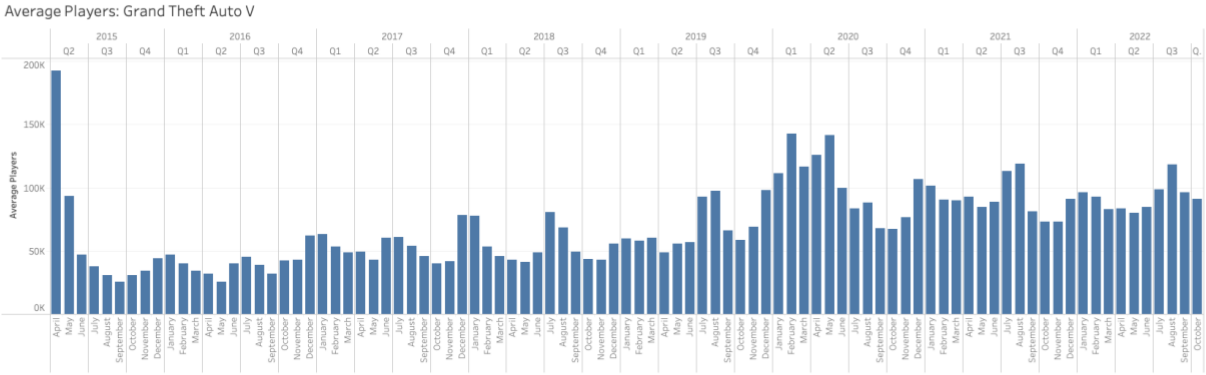

Figure 4: Average Players – Player Engagement

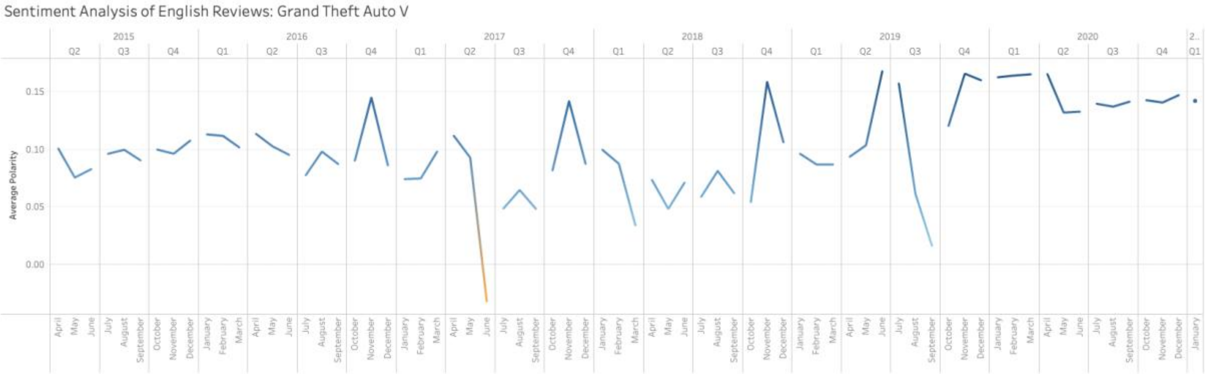

Figure 5: Review Sentiment

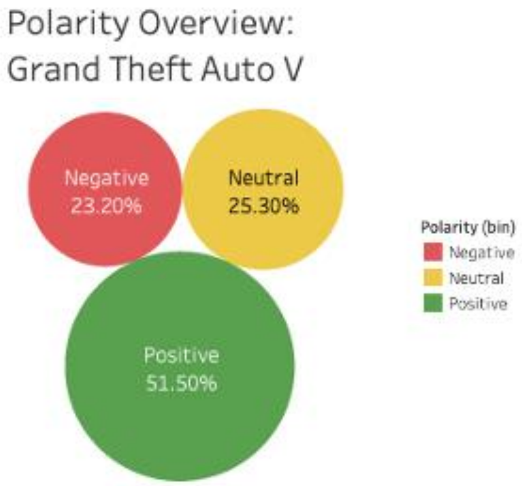

Figure 6: Polarity Overview - English Reviews

Choice of chart types

I want to show the relationship between review sentiment and player engagement over time. Player engagement decreases as the review sentiment goes in the negative direction. As I am correlating average players in Figure 4 with polarity in Figure 5 over time, a bar chart and a line chart effectively communicates the increase and decrease in the quantities. Bar and line charts are good at presenting categorical data that ordinary people easily understand. Lastly, the bubble chart in Figure 6 lets the audience quickly understand the overall proportion of different sentiment polarities for a specific game.

The data types of the visualised attributes are as follows. The polarity overview is a transformed view of the Sentiment Analysis line chart.

Attribute	**Data type**
Average players	Number
Polarity sentiment score	Floating point number
Polarity overview	Categorical

The polarity overview chart has the following tableau filter. The Sentiment Analysis line chart has visualised the average Polarity Converted value.

```
IF [Polarity Converted] < 0 THEN "Negative"
ELSEIF [Polarity Converted] > 0 THEN "Positive"
ELSE "Neutral" END
```

Design choices

I designed an interactive dashboard on Tableau with all three charts in one view with a filter that lets the audience pick and choose a game, as shown in Figure 7.

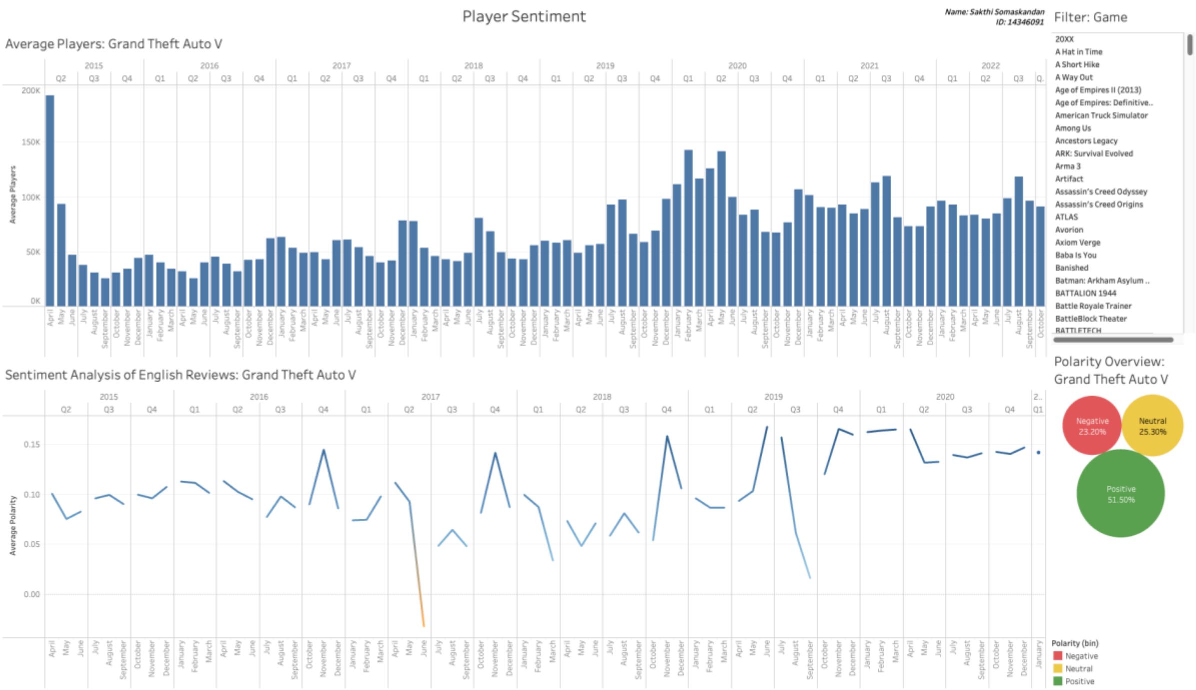

Figure 7: Player Sentiment Dashboard

The reasons for my design decisions:

- I structured the dashboard this way as it would be natural to compare and contrast the two charts in the centre and use the smaller right-hand side bubble chart as supplementary information.
- My choice of colour for the Polarity Overview chart is deliberately red, green and yellow as most people readily associate green with positive, red with negative and yellow with neutral (traffic light analogy). As the Polarity Overview chart is easily interpretable, the audience can focus on the main two charts, in the dashboard's centre.
- My choice of colour for the Sentiment Analysis chart is also deliberately made deep orange as it falls in the negative direction and deep navy as it rises in the positive direction. Deep orange is close to red, indicating an undesired outcome, and navy is a pleasant and neutral colour showing the desired result.
- I have used a line chart to display the polarity fluctuations to reduce the cognitive load on the audience when comparing the two graphs, as it would be arduous to compare two bar charts or two line charts.

Interactivity

I have the data for 300+ Steam games. Hence, I have provided a filter on the top right-hand side of the dashboard to allow the audience to interact and analyse the corresponding results for a specific game.

This aids in telling my story as we can compare the results across many different game titles. I understand that the story I want to convey is only valid when the hypothesis applies to many samples (games). Therefore, avoiding any misleading information propagating to the audience.

Interactivity is demonstrated in the screencast video.

Tools and libraries used

The tool that I used to create the visualisations is Tableau [3]. Python was used to gather data from Steam using one of their API endpoints and to clean the acquired data to aid in the visualisations. The python libraries that I used are:

- Matplotlib
- TextBlob
- Pandas
- Requests
- Json
- Time

Conclusion

The outcome of my visualisation successfully illustrates the relationship between review sentiment and player engagement. The relationship is proportional for most games, i.e., as the review sentiment increases (goes in the positive direction) the player engagement/game popularity increases. It is important to note that the sentiment scores were computed only for English reviews, and I filtered out other language reviews. As a result, the relationship proportionality may only convey accurately for some games, such as those with a high percentage of English reviews. However, the proportional relationship will be more visible when all language reviews are processed and given a sentiment score.

The biggest problem with the data was that the reviews were in textual format and the other columns in the dataset were straightforward for a good and insightful visualisation project. I spent most of my time thinking of ways to expand upon the obtained dataset from Kaggle – which I did by gathering player statistics and performing sentiment analysis on the review text so that the visualisation is accurate and valuable to the audience.

In terms of improvements, I wanted to show one chart with three axes – timeline on the `x-axis`, average players on the left-hand side `y-axis` and average polarity on the right-hand side `y-axis`. The two `y-axis` scales would be different as average players is in the range `[0, 200000]` and average polarity is in the range `[-1.0, 1.0]`. I tried to get it on Tableau but got an error: `Error Code: 73F9639A` (the full stack trace can be found on GitHub). A single chart would greatly benefit the audience by reducing the cognitive load and effectively communicating my story in a much smaller time frame.

Datasets

The datasets used to create the visualisations are available to download here. The submission zip file includes the Tableau file and the other relevant Jupyter notebooks. The video file of the screencast can also be accessed via the same link provided above.

References

[1] M. M, “Steam Reviews Dataset 2021 | Kaggle,” 2020.

https://www.kaggle.com/datasets/najzeko/steam-reviews-2021 (accessed Nov. 26, 2022).

[2] S. Loria, “TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation,” 2020. https://textblob.readthedocs.io/en/dev/#textblob-simplified-text-processing (accessed Nov. 26, 2022).

[3] “Business Intelligence and Analytics Software.” https://www.tableau.com/ (accessed Nov. 26, 2022).