STEAM DATA ANALYSIS

Marie-Christin Häge
Student
THU
Ulm, Baden-Württemberg, 89077
haege@mail.hs-ulm.de

Simon Mensch
Student
THU
Schwendi, Baden-Württemberg, 88477
smensch@mail.hs-ulm.de

Niklas Fromm
Student
THU
Weißenhorn, Bavaria, 89264
fromm@mail.hs-ulm.de

Stefan Eitel
Student
THU
Ulm, Baden-Württemberg, 89079
eitel@mail.hs-ulm.de

Sunho Ji
Student
THU
Ulm, Baden-Württemberg, 89073
ji@mail.hs-ulm.de

ABSTRACT

Due to the increasing number of gamers caused by the COVID-19 pandemic, this paper aims to analyze the behavior of the worldwide gamer community, as it has become a topic with higher importance. Therefore, we chose Steam as representative, which is a game platform offering a wide range of games and genres. In the paper questions regarding genres, players' locations and playtimes were answered with data collected from Steam.

Keywords: steam, game, genre, playtime, location, covid-19, data analysis, sql

1. INTRODUCTION

Due to the COVID-19 pandemic and its impact on the gaming industry, the idea of this paper is to have a closer look on the behavior of gamers worldwide, regarding the players' locations, game genres and playtimes.

The approach was to firstly define the questions to be answered and afterwards collecting the necessary data.

This paper attempts to analyze data from game users to discover insights for commercial purposes.

2. DISCUSSED QUESTIONS

The three main topics, which are supposed to be analyzed, are game genre, player's location and playtime. Therefore, the following questions will be answered:

Question 1a – "What are the overall most played genres regarding the playtime?"

Question 1b – "Regarding question 1a, what are the differences between countries of different continents?"

Question 2 – "What genres are released the most in recent months and years?"

Question 3a – "What are the most played games regarding the average playtime?"

Question 3b – "What are the most played games regarding the number of players?"

Question 4a – "What are the countries having the overall highest average playtime?"

Question 4b – "What are the countries having the overall highest amount of purchased games?"

3. DATASETS

In general, data has to be collected and then must be structured appropriately and refined to be used for specific purpose. By doing these tasks, structured data can be used to efficiently analyze large amounts of data and obtain meaningful results.

3.1 FINDING SUITABLE DATA

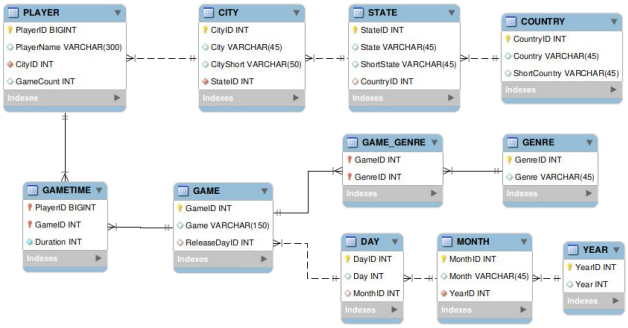
First, the data was searched on Kaggle, which is an online platform offering datasets to different topics. The data set imported from Kaggle did not meet the needs for the analysis. Therefore, the approach was to mine data by using a Steam API. The issue here was that a lot of deleted accounts had been collected, where no data was obtained. Consequently, another API was used, where public account ids were collected by starting with a public account and iteration over the friend list. This was repeated recursively. The resulting list of account ids served as a basis for collecting the necessary data, containing information like games and their playtime, the associated genres and the location of the players. Then another problem occurred as some information was only in form of ids, so another API was used to substitute the ids with their actual values.

3.2 DATA ORGANIZATION

Suitable data does not mean that there is nothing more to do, the quality can still be bad, so before answering the defined questions sample checks were performed to prove that the data is realistic and thus usable. As a next step, all unnecessary data was deleted, and disturbing symbols were removed. Furthermore, the dates and genres were atomized. Afterwards, .csv files were created to be able to finally import the data into a SQL database using KNIME and phpMyAdmin.

4. DATA WAREHOUSE

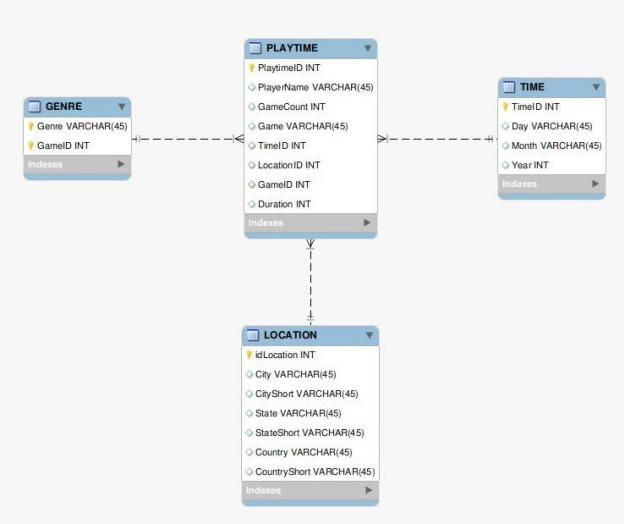
Every step was taken, so the structure, the data maintenance and the accuracy were guaranteed. At first the structure of the data warehouse was created, which consists of the table creations and the definition of keys and indices. The table creation aimed to result in a normalized snowflake schema with the three dimensions time, genre and location. The time dimension was split up into day, month and year and the location dimension was divided into city, state and country. Then the data was imported. Below the schema is displayed as EER diagram.



Each table contains a primary key named after the table. The location dimension is linked to the PLAYER table by a specific CityID. Even if some player data does not contain a city or state, the country is connected via state and city. The specific values for City and State are then null. The

GAMETIME table links the player with their purchased games (GAME table) and contains the playtime (Duration). The GAME_GENRE table serves as a linking table between the games and their genres, because of the existing n:m relationship. The release day is structured in the second dimension (time dimension) by splitting it into day, month and year. Therefore, the ReleaseDayID is linked to the DayID analogously as in the location dimension.

With SQL statements we created our data mart from the Snowflake Schema by joining and copying data into a new schema. We created the dimensions ”Time”, ”Location” and ”Genre” by 3 joining the outer tables with the inner tables. Furthermore, we reduced the number of countries to 12 and combined the fact tables into one table, so that we were able to query the data on the VM and to query it faster.



5. DATA PROCESSING

In order to answer the defined questions, SQL statements were created to extract specific .csv files, which contain data fitting exactly to the question. Consequently, six statements were executed, and their results transported to Excel and Power BI to finally analyze the data by creating pivot tables where it was necessary. Furthermore, these tools were used to visualize the results in different diagrams.

6. DATA INTERPRETATION

After everything was operated, the graphics can be used to answer the initial questions.

Question 1a – “What are the overall most played genres regarding the playtime?”

Ranking	Genre	Playtime in h(average)
1	Shooter	255,4
2	Soccer	153,6
3	Crafting	108,2
4	Military	90,0
5	Multiplayer	84,3
6	Clicker	78,6
7	Muiltiple Endings	78,1
8	MMORPG	76,5
9	Space Sim	75,8
10	Sandbox	72,4

To generate this table, only twelve representative countries were considered: Australia, Brazil, Canada, China, France, Germany, Japan, Korean Republic, Russia, Spain, UK and the USA.

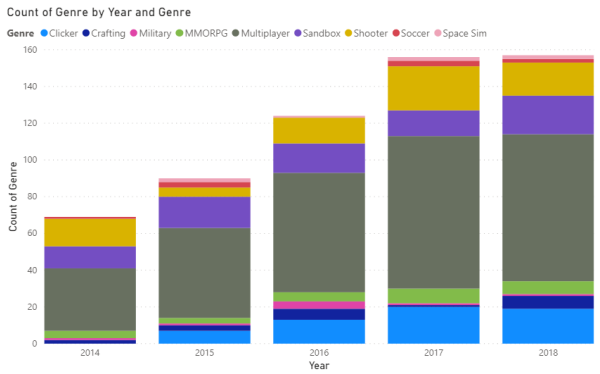
As predicted, shooters and soccer games are the most played genres. Interesting to see is that genres, which were expected to be more unpopular like clicker, military and spaces sim games, are also part of the ten most played genres.

Question 1b – “Regarding question 1a, what are the differences between countries of different continents?”

Ranking	Europe	Asia	Australia	America
1	Shooter	Multiple Endings	Shooter	Shooter
2	Soccer	Shooter	Soccer	Soccer
3	Military	Dating Sim	MMORPG	MMORPG
4	Multiplayer	Crafting	Crafting	Clicker

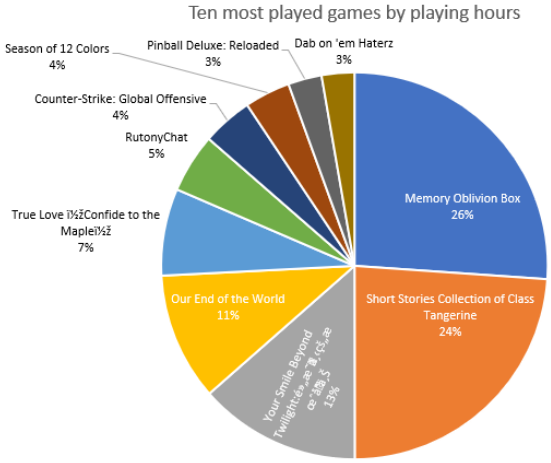
Comparing the countries of different continents by the most played genres, it is noticeable that except in Asia, shooters and soccer games are the most popular genres as it is worldwide. In Asia, the popularity of dating sim and multiple ending games is much higher than in the rest of the world. Especially in China, dating sim games are remarkably well liked. It is interesting to see, that Asian playing behavior is that different, which is maybe caused by their culture.

Question 2 – “What genres are released the most in recent months and years?”



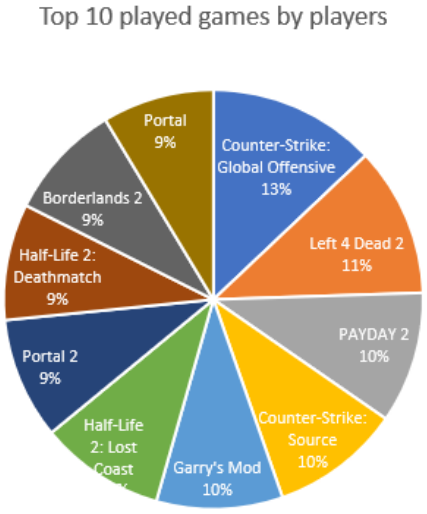
To answer the first part of the question regarding the differences between the months over 5 years, no clear pattern could be indicated. The expected outcome was to see that some genres are being released more frequently in certain months as e.g., more horror games are released in October. Furthermore, the second part was to compare the analyzed years. Multiplayer games are by far the most released genre. In general, no decisive differences can be noticed, except the fact that there is a steady increase of total game releases in the ten most played genres, defined in question 1a.

Question 3a – "What are the most played games regarding the average playtime?"



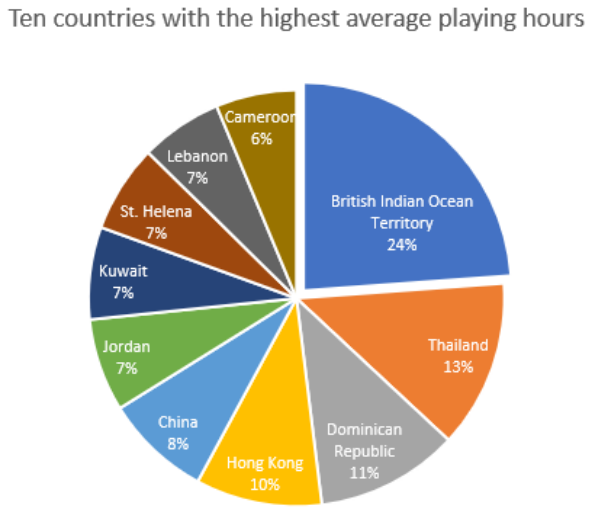
By looking at the pie chart, it can be indicated that that Memory Oblivion Box and Short Stories Collection of Class Tangerine are played the most. Those two games take up 50% of the playtime from the ten most played games. Interesting is that these games are produced by the same company and are not part of the most played and most popular genres. Therefore, games with the highest playtime don't necessarily have to be related to the most played genres.

Question 3b – "What are the most played games regarding the number of players?"



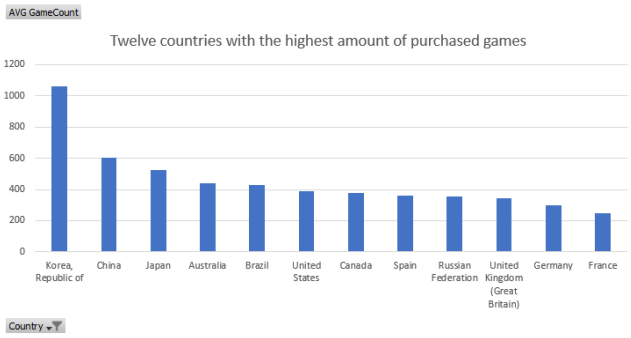
When answering the second part of question 3, the ten most played games are almost distributed evenly. The most played game regarding the number of players is Counter-Strike: Global Offensive with 13%, followed by Left 4 Dead 2 with 11%. Noticeable in comparison to the results from question 3a is that the distribution is split up evenly between the ten most popular games. Additionally, the most played game regarding the number of players, Counter-Strike: Global Offensive, only reaches rank six with four percent in question 3a. As a result, it can be concluded that there is a significant difference between the most played games by playtime or by owned games.

Question 4a – "What are the countries having the overall highest average playtime?"



For this question, no clear first place was expected. Therefore, it is highly interesting to see, that the British Indian Ocean Territory has an average playtime almost double as high as Thailand does. After the Dominican Republic and Hong Kong with each a bit less than Thailand, the following countries have no real difference in their average playing time.

Question 4b – "What are the countries having the overall highest amount of purchased games?"



The last question is like the previous one, but instead of considering the average playtime, the average amount of purchased games is being looked at. These countries are representatives for different continents. Asian countries, especially South Korea, are having the most games in their Steam library. After Australia being next, the American countries follow with a bit less games. The last part of the ranking is filled with European countries. For comparison, a South Korean player has in average about four times as many games then a German one. This is very fascinating and could be caused by the large amount of anime games being very popular in Asia. These anime games often have short story lines, so Asian players must buy many more games than German players, who are playing games with longer stories or even with open ending.

7. CONCLUSION

Overall, when analyzing the data, most of the results were as expected, as e.g., Counter-Strike: Global Offensive was the most played game regarding the number of players. Furthermore, the most popular genres were shooters and soccer games. On the other hand, some of the results were very surprising. E.g., Asian countries own a lot more games than the rest of the world. Noticeable was also that the most popular games in terms of players do not correlate with the most played games by playtime.

All in all, projects like this are interesting as they confirm certain assumptions but also surprise with new, unexpected results.