Games All Around the World

Daniel Glazer - 318282860

Valentin Volovik - 326814449

Table of Contents

1. Abstract

2. Introduction

Following the nested model learned in the course (3 question) that was inspired by Tamara Munzner (Munzner, 2009) and (Meyer, Sedlmair, & Munzner, 2012).

3. Background

The Valve Corporation a video game developer founded the Steam digital distribution service a decade and a half ago.

4. Methodology

In this chapter, there is a full documentation on how we collected the data on users and games from the Steam system and the data about countries. We also explain how we validate and analyze the data using visualization as will be shown further in the Data Analysis section.

4.1. Data Collection

Valve Corporation, the company that owns and operates Steam, provides a Steam Web API, for gathering information about users' profiles, friendships, game ownerships and playtimes, group memberships, and more. In the relatively new paper (O'Neill, Vaziripour, Wu, & Zappala, 2016) they use this very API to crawl 716 million games and more than 108 million Steam accounts, along with the information that is associated with each account.

Since their paper mostly focus on the user's relations like friendships and group memberships, we decided to focus on some interesting aspects other than the social aspect like Economy, Games and Gamer distribution.

The dataset comprises of our queries result on the dataset collected by (O'Neill, Vaziripour, Wu, & Zappala, 2016) on all Steam accounts available at the time of collection that specified the country they live in. We also obtained information countries worldwide with a GeoJSON dataset from Natural Earth. Both datasets mentioned above where collected in the past – the Steam dataset crawler collected data in 2013 – 2014 and the GeoJSON data is relevant to the year 2011.

Hence by definition this study is an observational study, or to be more precise a retrospective study – this means that while we can observe the data and establish associations / correlations we cannot establish causation between the explanatory and the response variable.

4.2. Data Validation

The data found in the Steam dataset was sampled manually to assure that accounts were associated with real users, both by randomly sampling hundreds of accounts and also by examining all accounts that exhibited extreme behaviors (the scrutiny includes examining their name, friends, and posts on their public profile). Needless to say, all the data collection was done by legal means – the data collected concerning user accounts is publicly accessible from player profiles, through both the Steam website and client.

Regarding the GeoJSON sourced from Natural Earth it is designed to meet the needs of production cartographers using a variety of software applications so we believe that the data is reliable.

4.3. Data Analysis

What?

Data and Dataset Types

Identifying the type of data is always the first step in the data analysis process.

The dataset is a combination of 2 datasets – GeoJSON dataset sourced from Natural Earth (that can be produced here https://geojson-maps.ash.ms/) and dataset that contains the results from queries (specified in the documentation to the derived data) on the Steam library dataset. The combination of those two results in a dataset in which there is both spatial data and relational data (tables), hence the type of the dataset is both relational and spatial. The dataset availability is static.

In this section only the variables of the derived data are shown (you can read about the variables of the raw data in the Steam website https://steam.internet.byu.edu/ or in (O'Neill, Vaziripour, Wu, & Zappala, 2016) paper.

For each **country**, we have:

Numerical:

Discrete:

gdp_md_est – an estimation of the country's GDP

money_spent – the amount of money spent by the country's players on games in the Steam library (in US Dollars)

pop_est – estimation of the population in the country

country_owners - the number of country's owners

country_active - the number of country's active users

avg_play_time - the country's average playtime (minutes)

num_casual_users - the number of country's casual users

num_moderate_users- the number of country's moderate users

num_excessive_users - the number of country's excessive users

for each X in range of 1 to 10 (for the 10 specific games selected)

gameXowners – the number of country's owners of game X

gameXactive_users - the number of country's active users of game X

gameXavg_play_time - the country's average playtime in game X(minutes)

gameXcasual_users - the number of country's casual users of game X

gameXmoderate_users - the number of country's moderate users of game X

gameXexcessive_users - the number of country's excessive users of game X

Categorical:

Regular Categorical:

continent – the continent's country

Ordinal:

economy – the country's economy group

income_grp - the country's income group

For each **game**, we have:

Numerical:

Discrete:

Appid – the game id in the Steam store

Is_Multiplayer – 1 if the game is multiplayer game, 0 otherwise

price – price payed to purchase a game

Required_Age – 0 if is suitable for all ages

Rating – the game rating (not for all games the ratings is specified)

Categorical:

Regular Categorical:

Genres – the game genres such as action, strategy etc.

Note: some of the properties are not specified but helped us to present the data to the user (such as country's name and iso_a2).

The next step in the data analysis process one would make is looking for relationships between variables.

A relationship between 2 variables could be either described as associated(dependent) or independent. Association can be Further described as either positive or negative.

Why?

So why would we even need a visualization of this dataset?

In general, any subset of statistical terms comes to mind can be computed in seconds and give as basic understanding of the dataset, however, this is only a general feeling of the data and will never give as the “full picture” (Anscombe's quartet is the most vivid example to this fact).

Specifically, in the Steam dataset …. //TODO

User tasks:

1. Present players distribution in various places(countries/continents)
2. Identify places with high percentage of addicts for specific game
3. Compare games’ addictiveness
4. Compare game popularity
5. Identify dependency between games properties
6. Identify correlations or similarities between game’s rating to the active players/ avg game playing time / owners
7. Identify correlations between GDP, money_spent and the economy of countries
8. Identify outliers related to country addictiveness

How?

To emphasize different aspects of the dataset, we divided the visualizations to 4 aspects:

- *Games*
- *Economy*
- *Countries*
- *Continents*

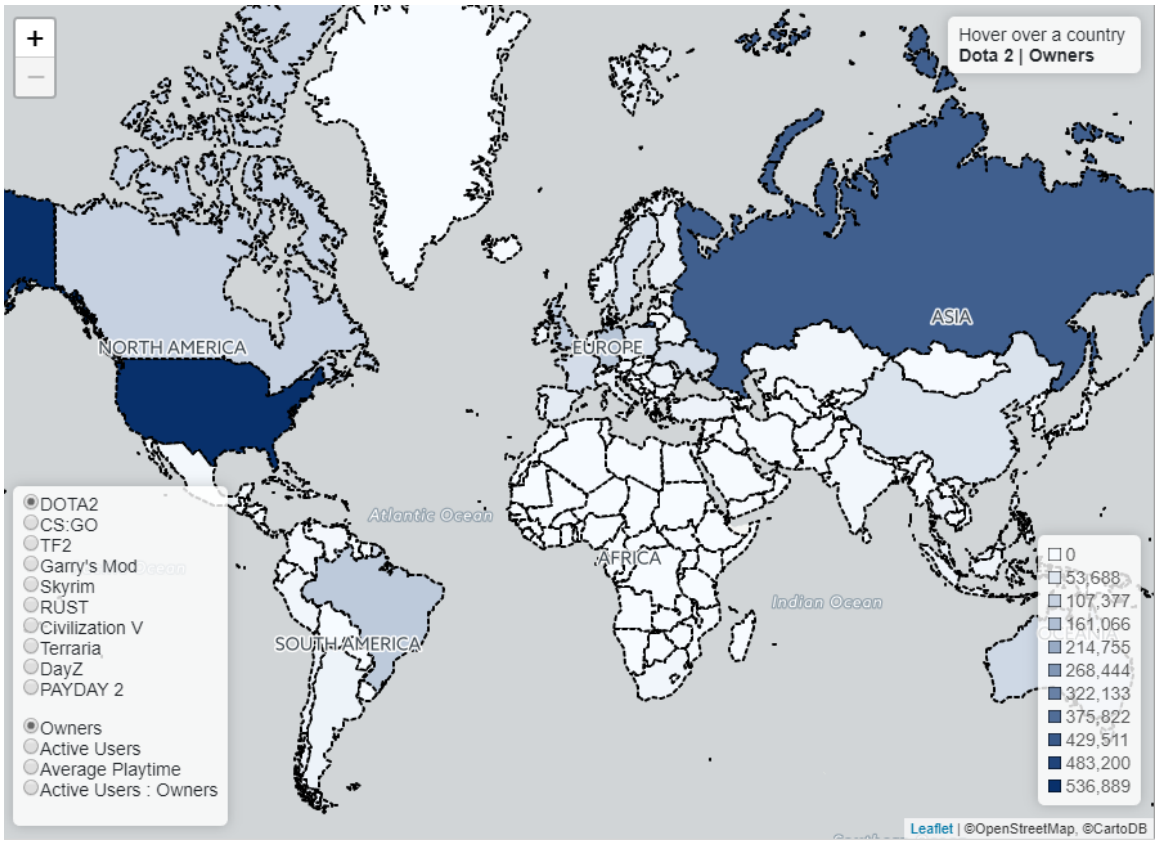
***We tried to prove in each aspect that even the simplest visualiz*ations can be the most powerful and best for the tasks in hand.**

Games

Per-game approach, where in each visualization the emphasis is on the game (one of the 10 selected), it’s gamers distribution and behavior.

Choropleth

To illustrate worldwide distribution of the players for the games that were chosen, choropleth was used with single hue progression. Leaflet map engine, along with CartoDB map provider for the labels, is used, providing the user option to explore the map. Hovering over the country shows the corresponding number / time for the chosen property. Legend in the bottom right corner serves dual purpose – both as a legend but also as a scale for the chosen property, as the darkest color is the property worldwide maximum for a chosen game.



Combination of game name and the property reapplies the choropleth with the chosen combination. In the example 'DOTA2' & 'Owners' indicates the number of owners of the game.
'Average playtime' was calculated for the 'active players' (users who were active in the timespan of 2 weeks at the moment of the data retrieval by the crawler that we derived the data from).

evaluation

This visualization saves **time** in**:**

- Finding extremes
- Finding anomalies
- Retrieving value

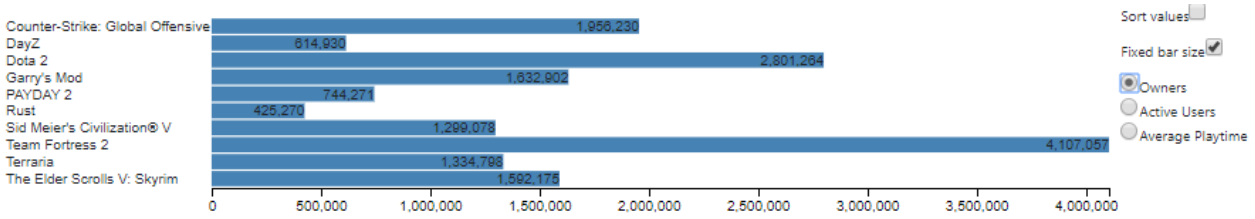
This visualization offers **insight** by giving the user the ability to find spatial trends in the map.

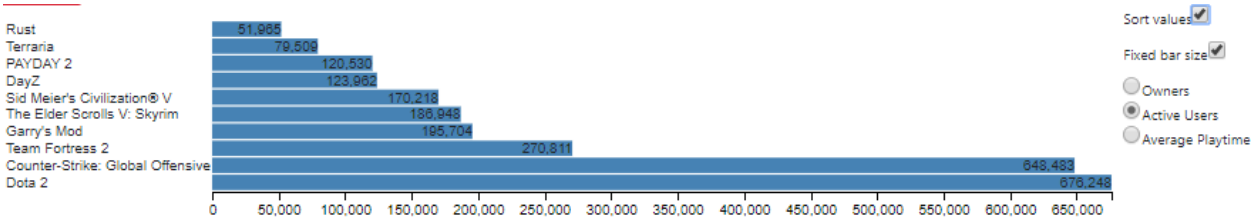
It emphasizes that some countries /spatial region have preference to specific games, even though the extreme stays for the most part the same country.

Bar Chart

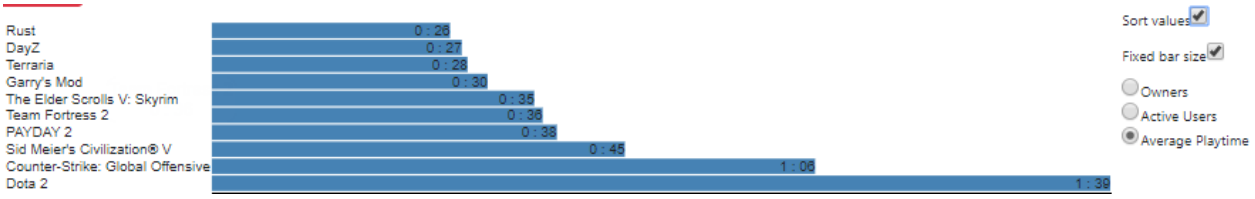
To illustrate worldwide how players are distributed for the 10 games, bar chart was used with single color, utilizing the bar sizes to indicate the difference.

The games aligned in the natural lexicographical order, with the option to sort the games according to the chosen property.





Average playtime was calculated for the total worldwide ‘active players’



evaluation

This visualization saves **time** in**:**

- Sorting the requested values
- Finding extremes / anomalies
- Characterizing distribution

This visualization offers **insight** by giving the user the ability to compare values in one screen.

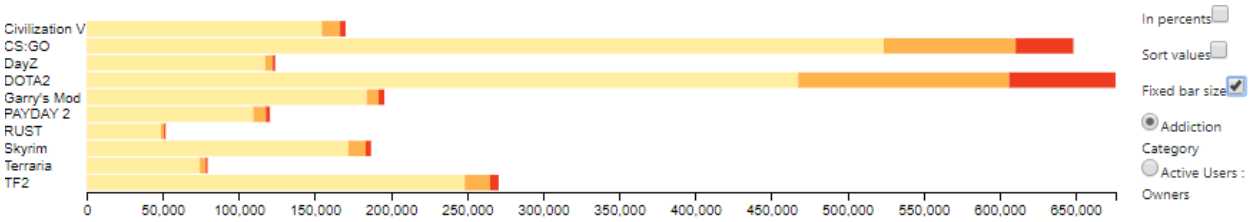
Stacked Bar Chart

As a direct follow-up to Bar Chart - Stacked Bar Chart allows to see the ratio of active players to owners and how can players be categorized by playtime hours.

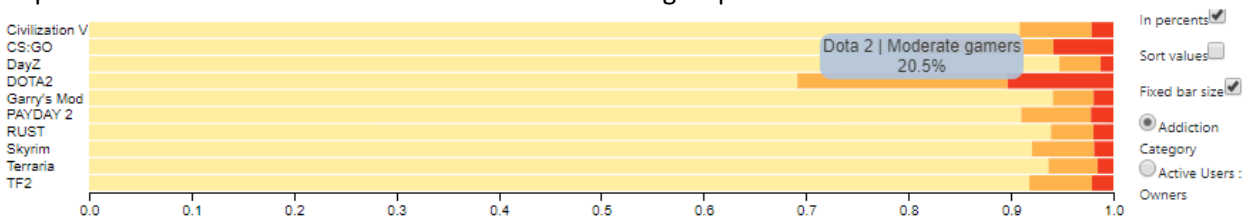
We divided the player's playtime to 3 categories:

- Casual : < 2 hours a day
- Moderate : 2-4 hours a day
- Excessive: > 4 hours a day

The chart can represented as raw number value or

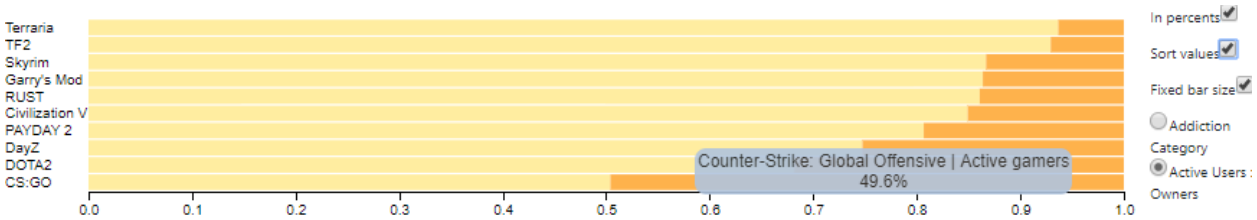


In percents where its easier to see the ratio between 3 groups



Alternatively, the view can be switched to illustrate the ratio between Active gamers and Non-active gamers

- Active Gamers : played this game in the last 2 week period
- Owners (Non-active) : haven't played this game in the last 2 week period



evaluation

This visualization saves **time** in**:**

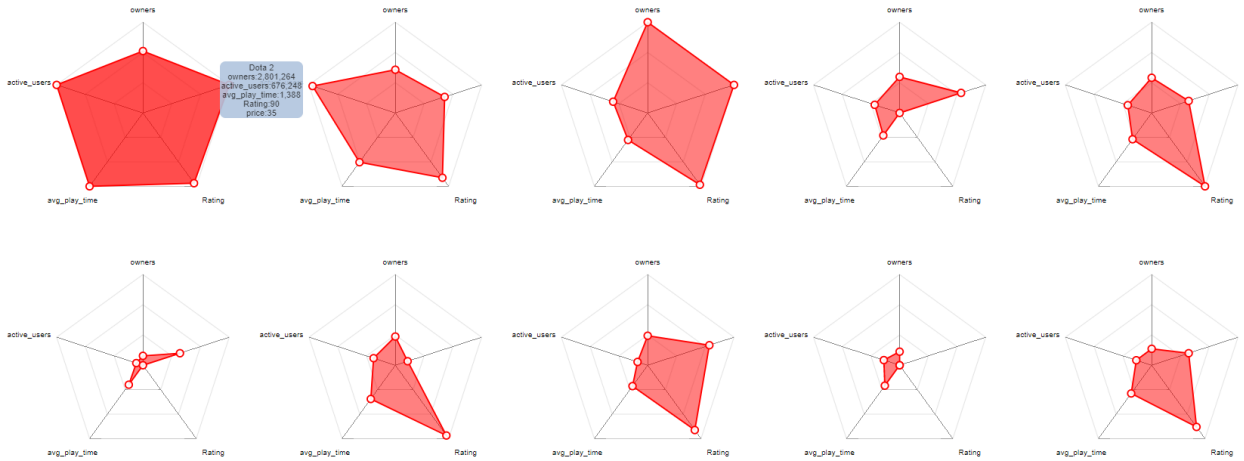
- Sorting the requested values
- Proportional sorting addictiveness of games
- Finding extremes / anomalies
- Characterizing distribution

This visualization offers **insight** by giving the user the ability to compare all the values and show proportion in one screen.

Radial Axis

To visualize the dependency of the game properties we choose to use the radar chart visualization.

The game properties selected are: number of owner/active users, average play time , rating and price.



Note: the price axis was reversed meaning the most expensive game is in the middle of the polygon.

evaluation

This visualization saves **time** in**:**

- Comparing derived values
- Finding extremes / anomalies
- Find correlation between game attributes (symmetrical shapes)

This visualization offers **insight** by giving the user the ability to see dependencies: - as the price goes down its popularity goes up (owners, active players and rating). Popular games has higher average playtime.

Economy

Line graph

In this simple visualization, we can see the positive correlation between GDP and Gamers' money expenditure on Steam games.

X axis corresponds to countries GDP.

Y axis corresponds to amount of money spent by Steam users in country.

The axis were scaled exponentially (power of 0.5 [x] and 0.4 [y]).

Color channel is used in three different division of countries to groups, according to :
Income Level / Economy Level / Continent.

evaluation

This visualization saves **time** in**:**

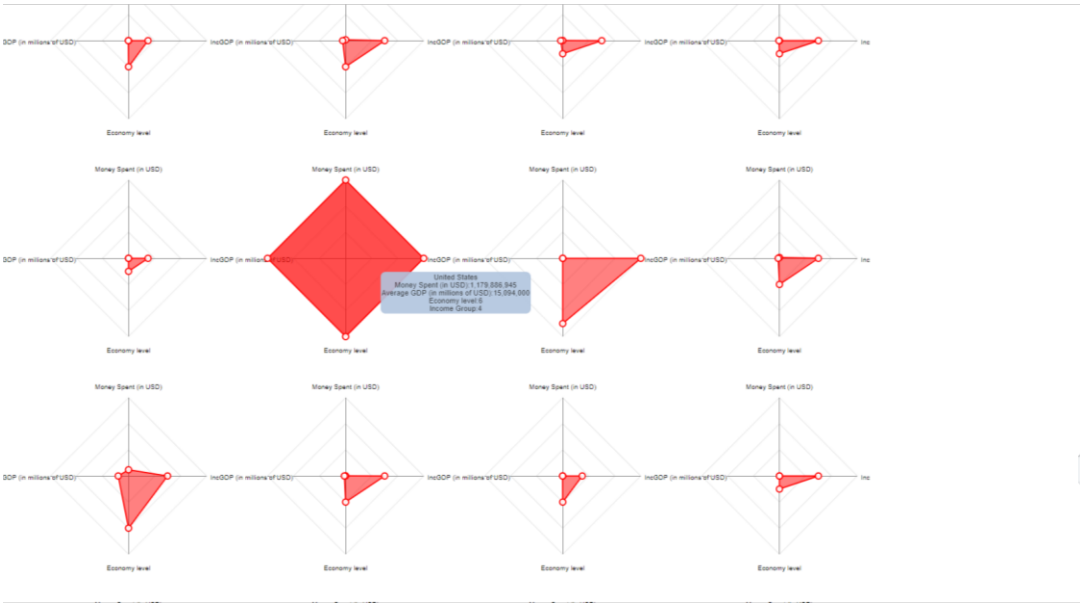
- Comparing derived values
- Finding extremes / anomalies
- Find correlation between country attributes (GDP / income group / economy level / money spent)
- Identify trends

This visualization offers **insight** on the trends in our data:

- Gamers in countries with higher economy level / income group compare to countries with the very same GDP are more likely to spent more money in the Steam store.
- Countries in the one continent compared to another country in a different, poorer continent with the same GDP value spent more money in the Steam store.

Radial Axis

In this visualization, we can see the positive correlation between country GDP /money expenditure/ economy level / income group on Steam games.



Note: The income group / economy level axis was reversed meaning the lowest income / economy group is in the middle of the polygon.

evaluation

This visualization saves **time** in**:**

- Comparing derived values
- Finding extremes / anomalies
- Find correlation between country attributes (symmetrical shapes)

This visualization offers **insight** by giving the user the ability to see dependencies:

- as the GDP / income group / economy level is higher so does the money expenditure in the Steam store.

Parallel Coordinates

In this visualization, we can see the positive correlation between GDP and Gamers' money expenditure on Steam games.

1st axis corresponds to Economy Level 0 – 6, where 6 identifies countries with the highest Economy level.

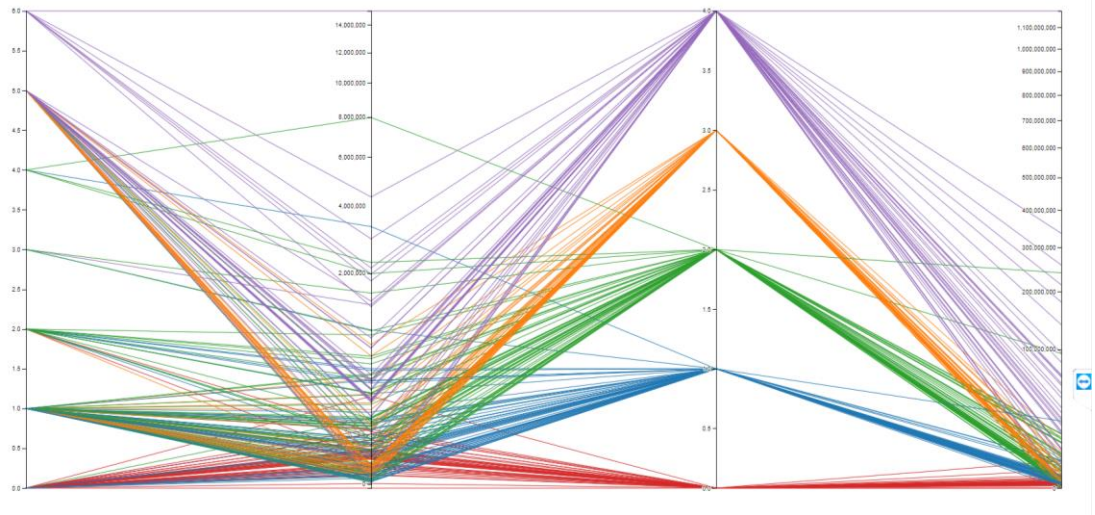
2nd axis corresponds to GDP estimate of the country.

3rd axis corresponds to Income Group to which the country belongs 0 – 4, where 4 identifies countries with highest income.

4th axis corresponds to the Amount of money spent by Steam Users.

2nd and 4th axis were scaled exponentially (power of 0.5, 0.4).

Color channel is used to identify countries with different Income Level.



evaluation

This visualization saves **time** in**:**

- Comparing derived values
- Finding extremes / anomalies
- Find correlation between country attributes (GDP / income group / economy level / money spent)
- Identify trends

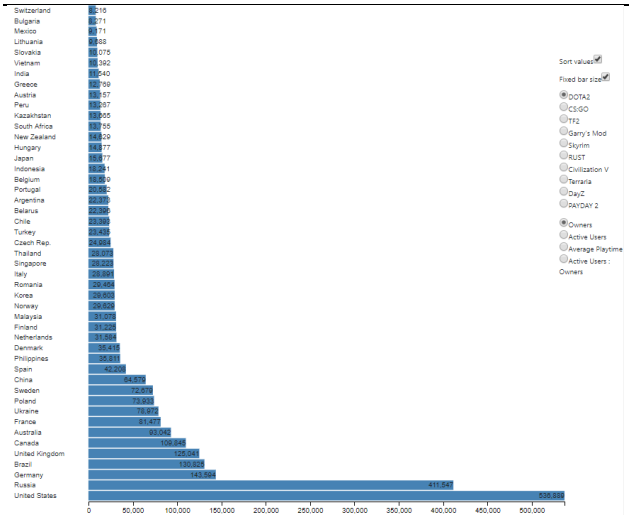
This visualization offers **insight** on the trends in our data:

- Gamers in countries with higher economy level / income group compare to countries with the very same GDP are more likely to spent more money in the Steam store.
- Countries in the one continent compared to another country in a different, poorer continent with the same GDP value spent more money in the Steam store.

Countries

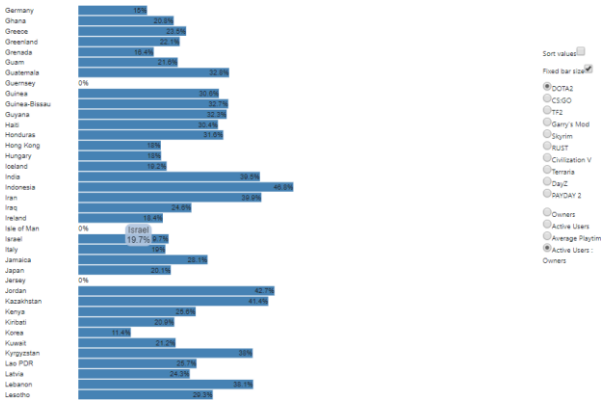
In this section, the approach is to look at the countries perspective: which countries boast the most gamers, and the biggest percentage

Bar Chart



Left: Sorted by value

Right: Lexicographical order



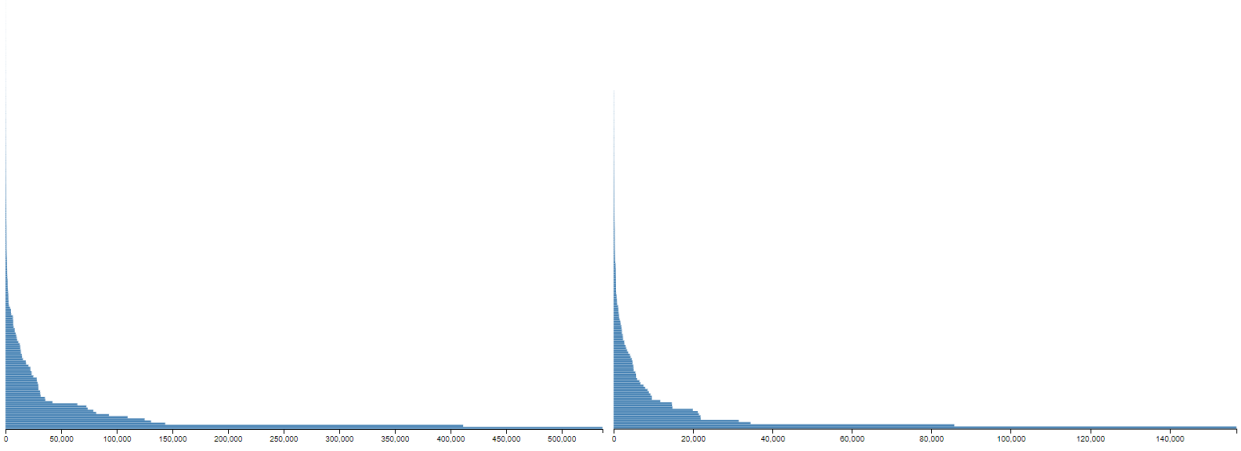
Along with the choropleth this chart provides simple distribution graph, where Countries are compared one along the other by specific-game-property combination

X axis is linear and always scales to the maximum value of property possible. (For example: “DOTA2”’s biggest base of participating gamers is USA and its value is 536,889, while “CS:GO” max value is 436,736)

Y axis size is controlled by a fixed bar height control; unchecking it allows to the exponential nature of the distribution, both at owners and active gamers

Left: DOTA2 owners

Right: DOTA2 active users



evaluation

This visualization saves **time** in**:**

- Sorting the requested values
- Finding extremes / anomalies
- Characterizing distribution
- Comparing countries values in different properties

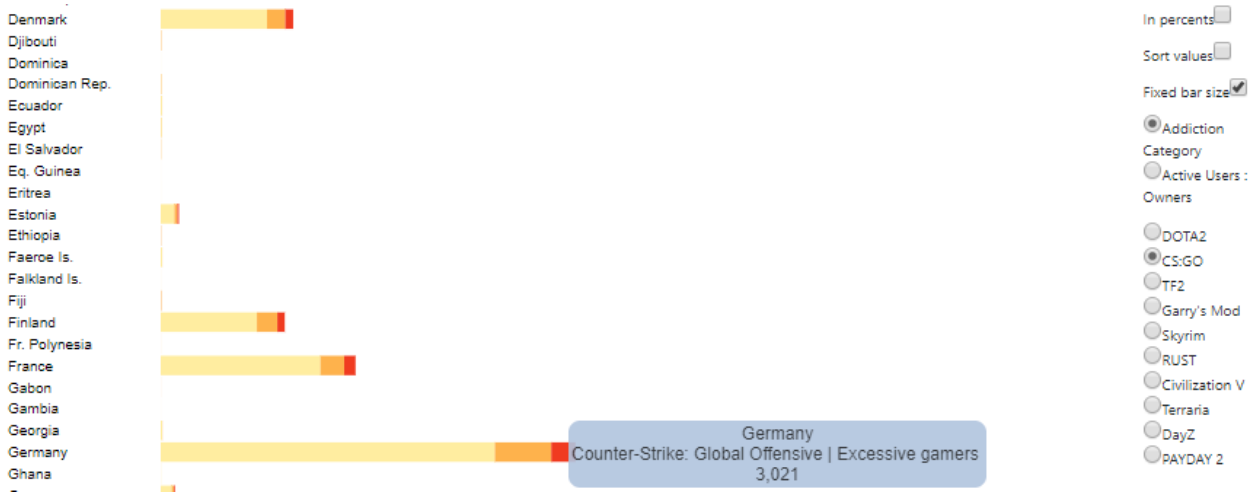
This visualization offers **insight** by giving the user the ability to compare values in one screen.

Stacked Bar Chart

Like in Games Stacked Bar Chart, in Countries Bar Chart there are modes:

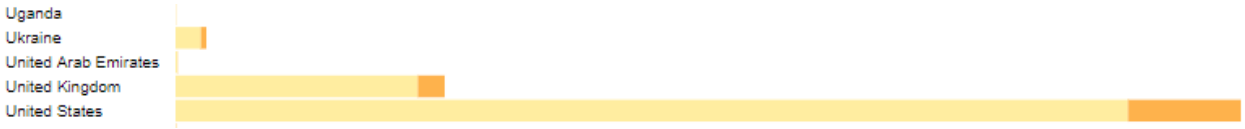
Players' playtime categories:

Casual : < 2 hours a day
Moderate : 2-4 hours a day
Excessive: > 4 hours a day

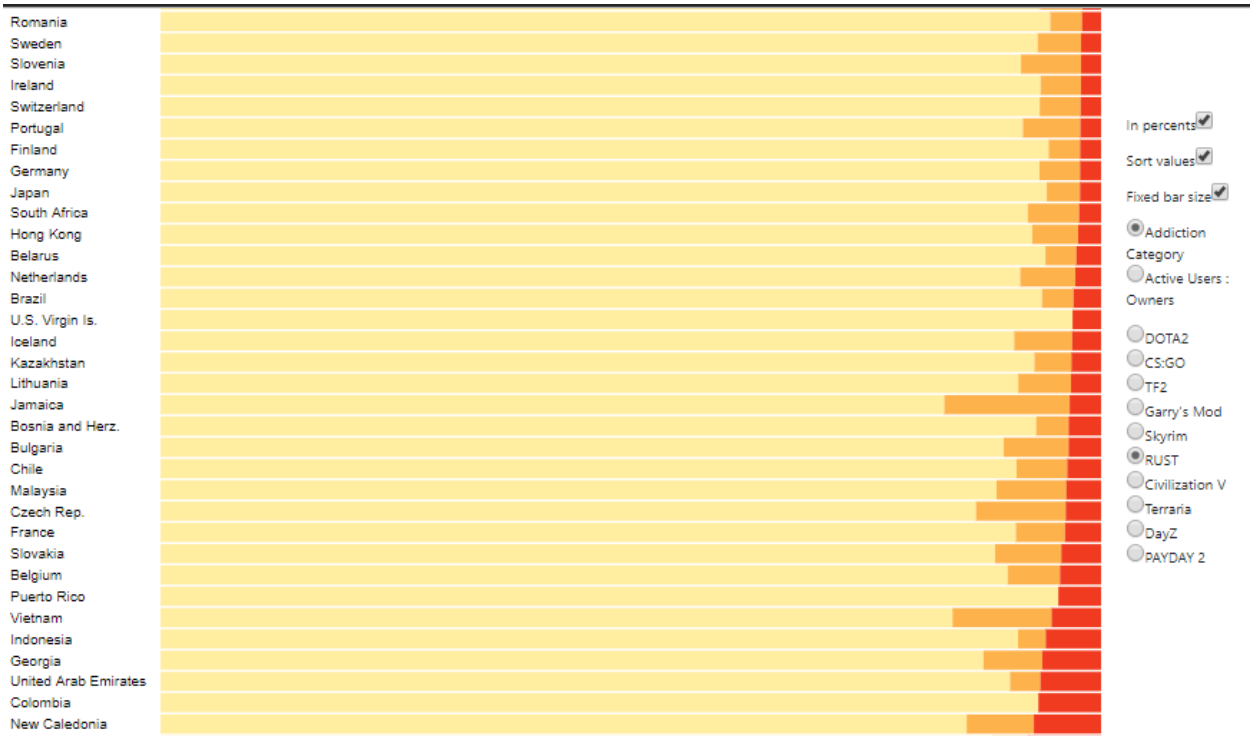


Per-country view active gamers to non-active ratio:

Active Gamers : played this game in the last 2 week period
Owners (Non-active) : haven't played this game in the last 2 week period



The view can be toggled to percentage view:



This visualization saves **time** in**:**

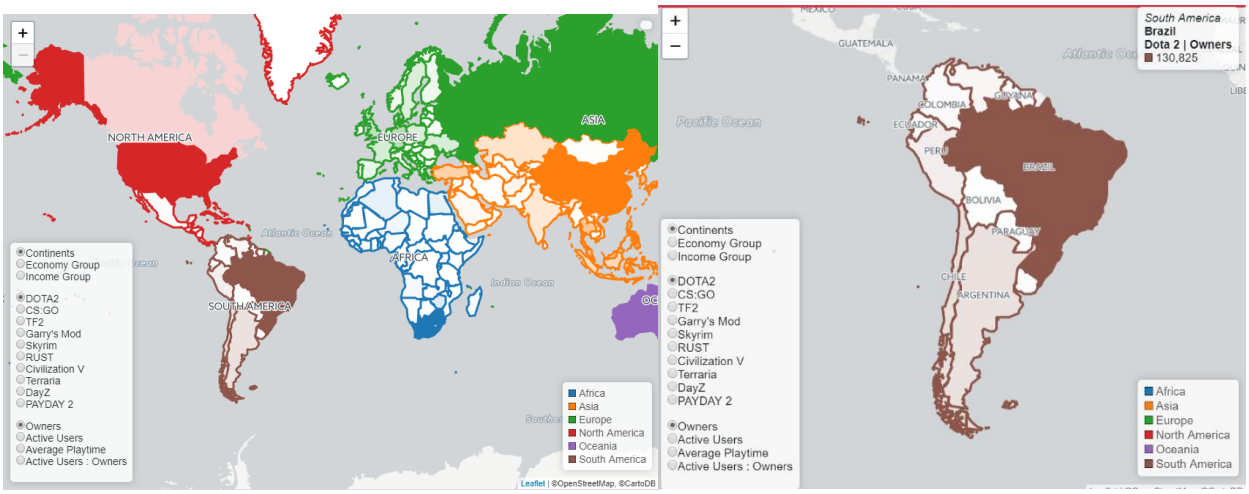
- Sorting the requested values
- Finding extremes / anomalies in the form of addictive countries
- Characterizing distribution of game active players in each country
- Comparing countries distribution of game active players

This visualization offers **insight** by giving the user the ability to compare countries addiction to specific game and also the ratio active: owners .

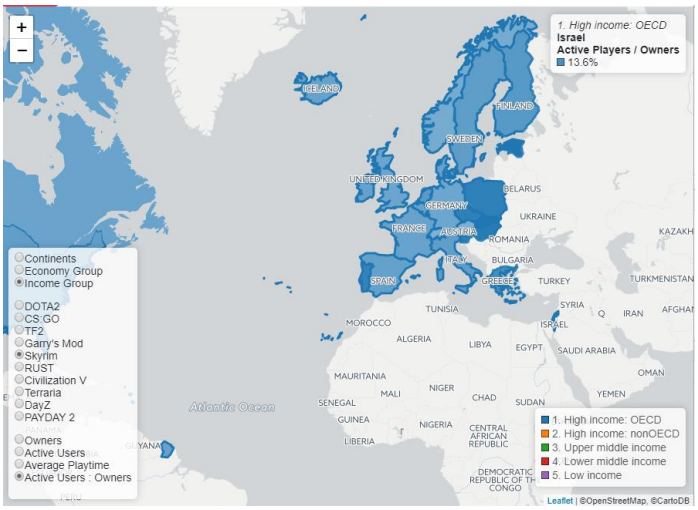
Continents

Choropleth

As can be seen in the previous chart – country like USA and its users' participation "dwarfs" all the other countries, per-continent / per-economy group division and choropleth provides geographical specific visualization



Left: Initial Right: on hover, South America



One of features of the GEOJson, is the option to divide the countries (with nesting) to different economy / income groups. This allows to observe distribution / percentage in this groups as well in the light of income selection.

evaluation

This visualization saves **time** in**:**

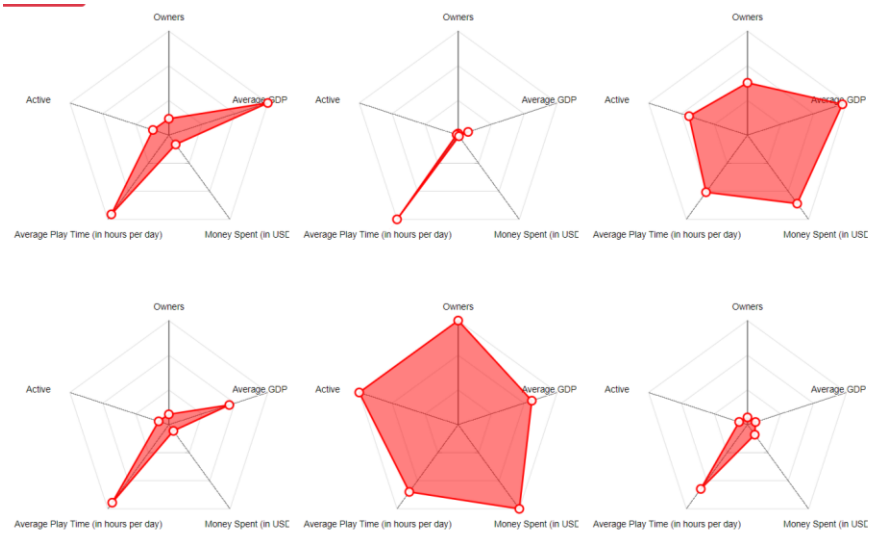
- Comparing derived values

- Finding extremes / anomalies in different groups
- Find correlation between country attributes (income group / economy)
- Identify local trends

This visualization offers **insight** on the trends in our data:

- In each continent, there is always 1 country, which property values are orders of magnitude bigger than the rest, but the county with the extreme values may be different per game on the same continent (indicating again the existence of game preference to locale)
- Countries in the one continent compared to another country in a different, poorer continents with the same GDP value spent more money in the Steam store.

Radial Axis



Note: the price axis was reversed meaning the most expensive game is in the middle of the polygon.

evaluation

This visualization saves **time** in**:**

- Comparing derived values
- Finding extremes / anomalies
- Find correlation between game attributes (symmetrical shapes)

This visualization offers **insight** by giving the user the ability to see dependencies: - as the price goes down its popularity goes up (owners, active players and rating). Popular games has higher average playtime.

5. Evaluation

6. Conclusions

7. References

Meyer, M. D., Sedlmair, M., & Munzner, T. (2012). *The four-level nested model revisited: blocks and guidelines*. Retrieved 8 17, 2017, from http://dl.acm.org/citation.cfm?id=2442587

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 921-928. Retrieved 8 18, 2017, from http://dl.acm.org/citation.cfm?id=1639181

O'Neill, M., Vaziripour, E., Wu, J., & Zappala, D. (2016). *Condensing Steam: Distilling the Diversity of Gamer Behavior*. Retrieved 8 22, 2017, from http://dblp.uni-trier.de/db/conf/imc/imc2016.html

8. Appendix

On the 10 games selected:

1. Dota 2 (Multiplayer Online Battle Arena)

Dota 2 is played in matches between two teams of five players. Dota 2 held the record for the game with the most concurrent users in Steam history, breaking its own record set in March of the same year. Simultaneous with this benchmark, the concurrent number of Dota 2 players in May 2013 outweighed the number of players for the rest of Steam's top ten most-played games combined.

2. Counter-Strike: Global Offensive (First Person Shooter)
   Counter-Strike (CS) is a series of multiplayer first-person shooter video games, in which teams of terrorists and counter-terrorists battle to, respectively, perpetrate an act of terror (bombing, hostage-taking) and prevent it (bomb defusal, hostage rescue). The series began on Windows in 1999 with the first version of Counter-Strike.
   As of August 2011, the Counter-Strike franchise has sold over 25 million units.
3. Team Fortress 2 (First Person Shooter | Multiplayer Online Battle Arena)
   Team Fortress 2 received critical acclaim for its art direction, gameplay, humor, and use of character in a multiplayer-only game. Valve continues to release new content, including maps, items and game modes, as well as community-made updates and contributed content. In June 2011, it became free-to-play, supported by microtransactions for in-game cosmetics.
4. Garry's Mod (Sandbox Physics)
   Garry's Mod (commonly abbreviated as GMod) is a sandbox physics game which was originally a mod for Half-Life 2, but was later made into a standalone release in 2006. As of January 2016, the game has sold 10 million copies.
5. Skyrim (Role Playing Game)
   The Elder Scrolls V: Skyrim is an open world action role-playing video game.
   During the first day of release, Steam showed over 230,000 people playing Skyrim concurrently. Skyrim had sold 30 million copies since its release in 2011.
6. PAYDAY 2 (cooperative first-person shooter)
7. Rust
8. 'Sid Meier's Civilization® V', (Grand Strategy Game)
9. 'Terraria' (2D action-adventure sandbox)
10. 'DayZ' (open world survival)