Information Visualization – G13-A – Checkpoint V Report

| Francisco Campaniço  IST  Lisboa, Portugal  francisco.campanico@tecnico.ulisboa.pt | João Rafael  IST  Caldas da Rainha, Portugal  joaolucasrafael@tecnico.ulisboa.pt | Rodrigo Oliveira  IST  Lisboa, Portugal  rodrigo.domingues.oliveira@tecnico.ulisboa.pt |
| --- | --- | --- |

# INTRODUCTION

Our project is related to electronic sports, more commonly referred to as Esports (eSports is a common way of spelling this as well, but it’s an incorrect one). Esports differ from “normal” sports because instead of being a primarily physical activity; while any sport will have a component of tactics behind it, there’s also a much higher requirement for physical prowess. The “electronic” component refers then to the fact that this kind of sport is related to playing video games in competitive tournaments, which will, of course, change the necessary requirements for success (there’s a lesser reliance on more “broad” physical conditions, but a higher one in hand-eye coordination, reflexes and tactics (and, depending on the game being played, additional requirements may include a good aim, or being able to manage several characters both effective and quickly, for example).

Depending on the game being played, the team sizes vary from one to usually five. In larger size teams, there’s a captain who will not only play in the team but also make the bigger tactical decisions, and a coach who isn’t an active player, but advises team members in both training and mid-game. Outside teams, there are always (usually two) *casters* commenting on the game, and a panel of commenters who, between games, give some opinions on team performance and the current *meta* (that is, what characters or weapons are most picked).

Esports have become a prevalent medium of entertainment in the past decade, as they combine the current generation’s rising interest in videogames with the already established interest in sports (for example in Portugal, where football fandom is prevalent). They combine the appeal of videogames with the stories that can come out of sports, such as the up-and-comers defeating big established teams, the newest star who rose to the top in their first year, the dismaying loss of a highly praised team in only the first games of a tournament. Our primary motivation is that we both enjoy playing videogames and watching Esports tournaments, as they provide great entertainment.

While it’s easy to find quite a lot of information for Esports, this information is either not organized in a particularly appealing fashion for quick assessment (one cannot conclude many correlations by browsing several player or team pages in wikis) or the only thing that’s able to be assessed is the games which award the highest prize pools, or the highest earning players. It’s not possible to correlate the number of players in a country with its more general statistics (such as, for example, urban population, which would be interesting to do in this case due to the electronic nature of Esports, meaning it requires access to the Internet).

## Tasks

Our approach allows the assessment of more interesting statistics related to Esports besides simpler information such as “highest earning teams”. In Checkpoint I, we decided that our visualization should have the following tasks:

* Analyzing how esports have evolved through the years;
* Searching for the countries with the highest player earnings;
* Identifying at what ages players earn the most;
* Comparing different teams;
* Comparing different games;
* Comparing at which months most tournaments were held.

Our visualization allows all these tasks. Moreover, it specifies some of them further, and adds some new ones:

* One can not only identify at what ages players earn the most, but one can also filter this information per the desired country;
* It’s possible to compare different games and teams through both monetary statistics (prize pool sum for games, earnings for teams) and number of tournaments hosted/played;
* One can correlate a country’s number of players or their earnings with some of the country’s statistics;
* It’s possible to see what teams won tournaments in each month, ranging from 2000 to 2017.

# RELATED WORK

We didn’t find much related work in the field of Esports. It’s common to find infographics (images displaying several interesting stats) such as [this one](https://www.reddit.com/r/DotA2/comments/6jxwcg/infographic_for_the_ti7_open_qualifiers/) for a particularly popular tournament, or “top 100” rankings for earnings or prize pools, such as the ones found on [*esportsearnings.com*](https://www.esportsearnings.com/), but we didn’t take an inspiration in any “in-depth” visualization for the theme. As we stated in the introduction, there weren’t visualizations that went beyond very game-specific statistics (such as the infographic above) or comparing prize pools.

# THE DATA

Our data was obtained from two websites’ APIs: [*esportsearnings.com*](https://www.esportsearnings.com/)for esports-related dataand [*worldbank.org*](http://www.worldbank.org/) for more general country statistics. The statistics for *WorldBank* included urban population, unemployment, GDP and education expenditure. Sadly, the education expenditure statistics were too incomplete for us to use them.

## Issues

The *EsportsEarnings* API didn’t display all the information available on the website, and since the data relied on user submissions, we were unable to use data for additional tasks. For example, while something such as the location where tournaments are hosted would be quite interesting, and the data was available in the API, due to it being user submitted there was too much variation in location format, such as “California, USA” (state, country) and “Greater Warsaw, Warsaw, Poland” (metro area, city, country), and we couldn’t use it.

Another problem with the API is that each user is allowed only one query every two seconds, and the queries are very limited. This proved particularly annoying when we decided to add some interactivity where you can see what teams won (and how many tournaments they won) in a month of the user’s choosing. To get all the tournaments (which we already had to do), we could only query the “100 most recent” with an offset which allowed us to move back to the first tournament in the website’s data (which took place in 1996, for the curious). Since the website had around 32 thousand tournaments, this required 320 queries, and if one does the math, 10 minutes to perform all necessary queries.

This isn’t a lot of course, but the thing is this query doesn’t provide information for who won the tournament. To obtain that kind of information, a different query is required, one for every tournament. This adds up to about 17(!) hours of leaving an automation script querying everything. We put it running on a Google server (thankfully IST provides free server uptime in these services) and we only remembered to check back a week after.

When trying to add functionality for selecting a country and seeing at what ages players earn the most in that country (we didn’t have that data when we delivered Checkpoint II, but we ended up wanting to add that feature), we also ran into the problem that Pentaho DI (the recommended program for data processing) doesn’t work well with nested JSON tables. The way we had this organized was that for each player, there was a sub-table with age values and what they earned for each age. We managed to make it work with the global data, but for each country, the solution was to just return this data for each country and… join it in a file by hand. This resulted in about three hours of heavily repetitive work (the factory scene in Charlie Chaplin’s “Modern Times” where he does extremely repetitive and mechanical actions came to mind) and a quite starved IST student. And the data we wanted, of course.

## Compromises

We had to make several compromises in the data shown as well. The aforementioned 32 thousand tournaments ranged from 1996 to 2018, but the 1996-1999 range only included about 10 tournaments, which wouldn’t be very interesting to display as only roughly one fifth of months in those four years would actually have a tournament. Aside from that, since 2018 isn’t over yet, it wouldn’t be right to display data related to that month as well, so we ended up cutting those years out, resulting in a 2000-2017 range.

Also related to tournaments, the data for what teams won tournaments in a selected month was sadly quite incomplete. However, we felt that this information, and the interactivity it provided to our visualization, was too interesting to pass on, and so we ended up leaving it as-is.

For the “earnings by age” statistic (what ages players earn the most), we also had incomplete data, as some players didn’t have specific age-related data. This data was not even acquired through an API and was instead scraped from each player’s *EsportsEarnings* age-related page. As above, the information and interactivity given is far too interesting to pass on, and we left it as-is.

Several data required the use of Python scripts for us to get a usable table. For example, we had to add some IDs that the *EsportsEarnings* API used to some tables that didn’t have them, as they were required for some interactivity actions.

## Resulting data

We ended up with a not-too-large 15 MB of data, including around 4000 players, 400 games, 700 teams and 32000 tournaments. Since there were a lot of attributes that we didn’t need in these tables (players’ nicknames and real names, and tournament locations for example), we ended up filtering quite a lot of data that was simply unnecessary for our visualization, which resulted in only slightly above 2 MB of data.

With d3.js, this proved to not have scalability issues: The framework is probably built to handle much larger amounts of data, so this didn’t really prove to be an issue. Our final data still included the same number of teams and games, but since we didn’t need individual players or tournaments, we were left with derived tables such as the number of tournaments in each month, or the number of players in a country.

Aside from the feature where one selects a country and see earnings/age distribution for that country, which required us to revisit data processing when we were implementing it in Checkpoint IV, and adding a few IDs to make some interactivity work, we didn’t feel the need to restructure much of the data that we had already acquired in Checkpoint II.

# VISUALIZATION

## Overall Description

Leave me here to keep the formatting, please.

## Rationale

Leave me here to keep the formatting, please.

## Potential

Leave me here to keep the formatting, please.

# IMPLEMENTATION DETAILS

Aside from d3.js, we used jQuery to ease some interaction such as tooltips.

While d3 does have quite a lot of features, the documentation is sub-par at times (consisting of a lot of disjointed GitHub repositories), so it took us some time to find decent scale types, such as the power scale, which ended up being more useful than a logarithmic scale simply because it also allows displaying different orders of magnitude effectively (unlike a linear scale) but it has a customizable exponent property, which solves the logarithmic scale problem where a lot oh higher values are near-indistinguishable, since one can set the exponent at a low value to allow both low and high orders of magnitude to have detail.

Another important function in d3 was the formatting. It’s possible to format numbers with quite a lot of customization, and we used that for very different things. We usually display large numbers with an SI prefix (k for thousands, M for millions, et cetera), but in certain situations we decided to display the whole number with comma separation (for example, $1,234,567). The formatting is also useful for displaying percentage and dollar signs without messing around with strings too much.

# CONCLUSION & FUTURE WORK

Leave me here to keep the formatting, please.