UNIVERSITY OF CALGARY

BTMA 431 Final Project Report

Group 3: Lincoln Baines, Arjan Birdi, Cooper Chung, James Zhou

Haskayne School of Business, University of Calgary

BTMA 431: Gathering, Wrangling, and Analyzing Data in R

Professor Meysam Fereidouni

December 11, 2022

Introduction

For our final project, we decided to study eSports. Specifically, we focused on three areas; prize pools of games, earnings of teams, and overall game popularity. Our motivation to conduct this study came from our shared interest in eSports. As a result of COVID moving our everyday lives to the virtual world, our need for entertainment shifted online as well. To many that are already familiar with competitive video games, eSports is nothing new. However, COVID was a major breakthrough for eSports as it was allowed to finally garner the attention of the general public. As eSports became a larger, more internationally recognized sporting category, we saw an opportunity to evaluate the current earnings landscape and growth.

We sought to answer the questions that prospective players and teams would ask when initially taking their first steps into becoming established in the eSports scene. Moreover, we would like to provide insight and data for already established teams and players on becoming more successful. We studied questions such as;

1. If you were to compete professionally in eSports, which game should you choose to make the most money?
    a. Which games have the lowest and highest average prize pool per event?
    b. Is there a relationship between the total amount of prize money for a game and the number of players?

2. Which team/organization has the highest amount of prize money won?
    a. For the team with the highest amount of prize money won, what is the percentage distribution of their total earnings from all games they participate in?
    b. Is there a relationship between the total amount of prize money won by a team and the number of tournaments the team participates in?

3. In the past year, which game has been the most popular in terms of viewership?
    a. How has covid affected the top viewed games and total viewership?

This data would also be helpful for other eSports analysts looking to gather some numbers and develop a better understanding of earnings in eSports. We would also like to cater to the general audience that consumes eSports content and help them pick the most

popular events to watch. With many newcomers, whether it be a new venture looking to pick up its first squad of players for a specific game, or a budding talent looking to make a name for themselves, we hope that our data and visualizations can assist teams and players alike in being successful.

Data Sources and Limitations

For our data sources, we decided to utilize two websites; esportsearnings.com and sullygnome.com. Firstly, we decided to use esportsearnings.com because they offered a free-to-use public API along with a simple, accessible HTML structure that made scraping data much easier for us. As for sullygnome.com, they provided us with premade data in CSV form, which made it incredibly simple for us to import into R and perform statistics. When utilizing sullygnome.com, we could change the date range of the CSV file to any possible range in the past 6 years. For consistency sake, we decided to use data from the past year on the popularity of games on twitch.tv. Overall we found that these two websites were a perfect fit for our project requirements.

However, this was not always the case for some other websites. We initially had other sources in mind, but upon trying to collect data from them, we found that they were surprisingly “anti-collection”. For example, our initial plan was to utilize the liquipedia.net API to gather data, but it turned out that the API was actually used for updating and editing the actual wiki itself, and it was not meant to be used to gather information. Furthermore, other websites such as escharts.com and twitchmetrics.com had their features locked behind a paid subscription. What this meant for us is that we were unable to access their API if they had one, and scraping data off of their websites was impossible as they intentionally returned an error to us, or there was no scrapable table in the HTML. While both websites contained interesting data that we could have used and would have made our project simpler to complete, without spending the necessary amount of money, we were ultimately unable to utilize their information. As a result of this limited availability of data, we needed to be specific with the questions we were asking and answering. Since eSports is also a

growing industry, we had a relatively low number of observations to work with, and as a result, outliers heavily skewed our data.

Methodology

For question one, data was obtained by scraping esportsearnings.com, specifically the “most prize money awarded” page in the “game” section of the website. This was done through the rvest library in R. Due to the unique structure of the top 5 games on the page, these had to be handled through different means. This required obtaining specific html elements through the use of xpath(). The numerical values were also parsed using parse_number() in the readr package. Once the data for the top 5 games was obtained, this was placed into a data frame. The remaining games were in an html table, which was easily parsed into a data frame using the readHTMLTable() function from the XML package. The two data frames were then combined, to obtain a complete set of information for the top 100 games. This was then exported to csv using the write.csv() function so that it could be imported into Power BI for analysis.

We also wanted to create a regression to determine whether there was a relationship between the total amount of prize money for a game and the number of players competing in that game. To do that, the lm() function was called with the data frame containing the information on the top 100 games, with it being asked to fit the “prize” column on the “players” column. For question two, we collected and visualized our data by utilizing a mixture of libraries, API calls, and HTML scraping methods. We utilized libraries such as httr and jsonlite to communicate with the esportsearnings.com API to retrieve raw data, and the rvest and XML libraries to scrape data we were interested in that was not included in any of their own functions. We also utilized ggplot2 to visualize our regression models.

First, we registered accounts on the esportsearnings website and obtained our own API keys. Second, we utilized one of their own API functions to obtain raw data regarding the highest-earning teams in eSports. We used a GET() function to download the page data,

converted the contents to JSON characters using the rawToChar() function, and stored the characters in a workable data frame using fromJSON().

Next, to find the distribution of the earnings of the top team, we scraped a page containing team earnings by game, as it was not a function included in the API. To accomplish this, we nested functions from the rvest and XML libraries such as getHTMLLinks(), htmlParse(), and GET() in order to retrieve the elements of the page of interest. Upon doing so, we identified the general pattern the URL followed when looking at specific teams, and realized we had enough data to automate the scraping of these web pages. We cleaned the existing data frame we generated from the API using gsub(), and assembled a list of links by pasting it onto the pattern of the URL using paste0(). We automated the collection by creating a function that intakes a URL, reads the HTML from it, identifies tables of data to collect, and creates a data frame from this table. Applying this function to the list of links created earlier, we harvested a list of 100 data frames, each containing data of interest for each team. We confirmed that our data was sorted from highest to lowest by ensuring that the first team in this data frame was still consistent with the information we knew from earlier, and broke down their earnings from there. The resulting data frame of the highest earning team was written to a CSV file, which we then imported into PowerBI in order to generate visualizations.

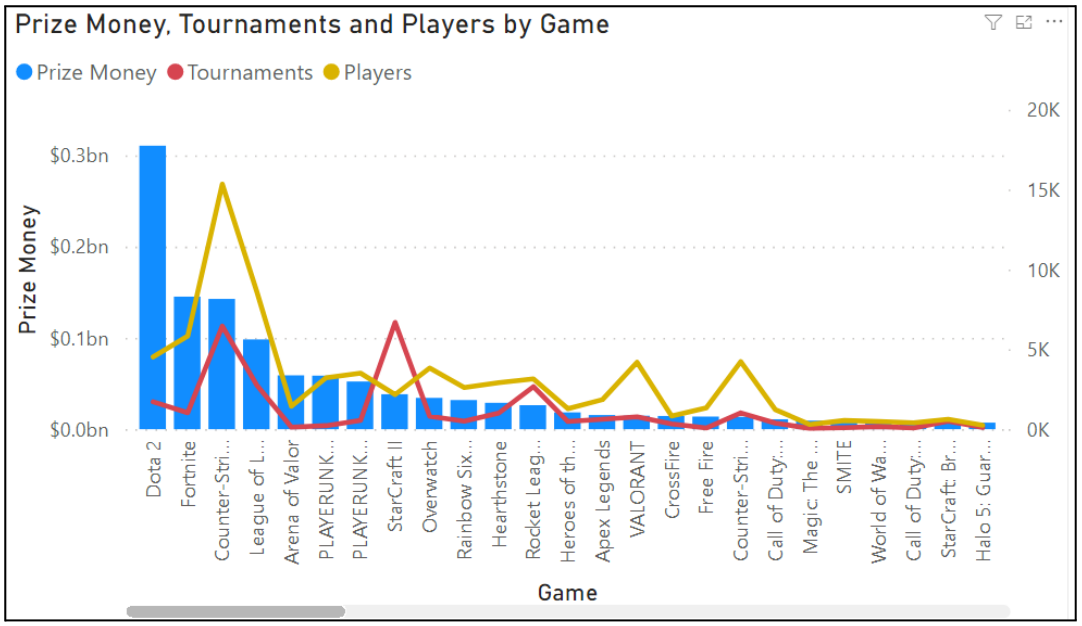
We also wanted to create a regression to find out if you could predict a team's winnings based on the number of tournaments they play in. In order to perform a regression, we first had to perform a logarithmic transformation to reduce heteroscedasticity because the scale of the data was so large. We created a linear regression model on the log of the total USD prize money of a team, using the log of the total number of tournaments the team has participated in as a predictor. To visualize this regression, we utilized the ggplot2 and scales libraries to create a graph.

In gathering data for question 3, we were able to download CSV files from sullygnome.com that contained relevant data. Once downloaded, we then proceeded to import the datasets into R Studio using the read.csv() base function. Importing it into R

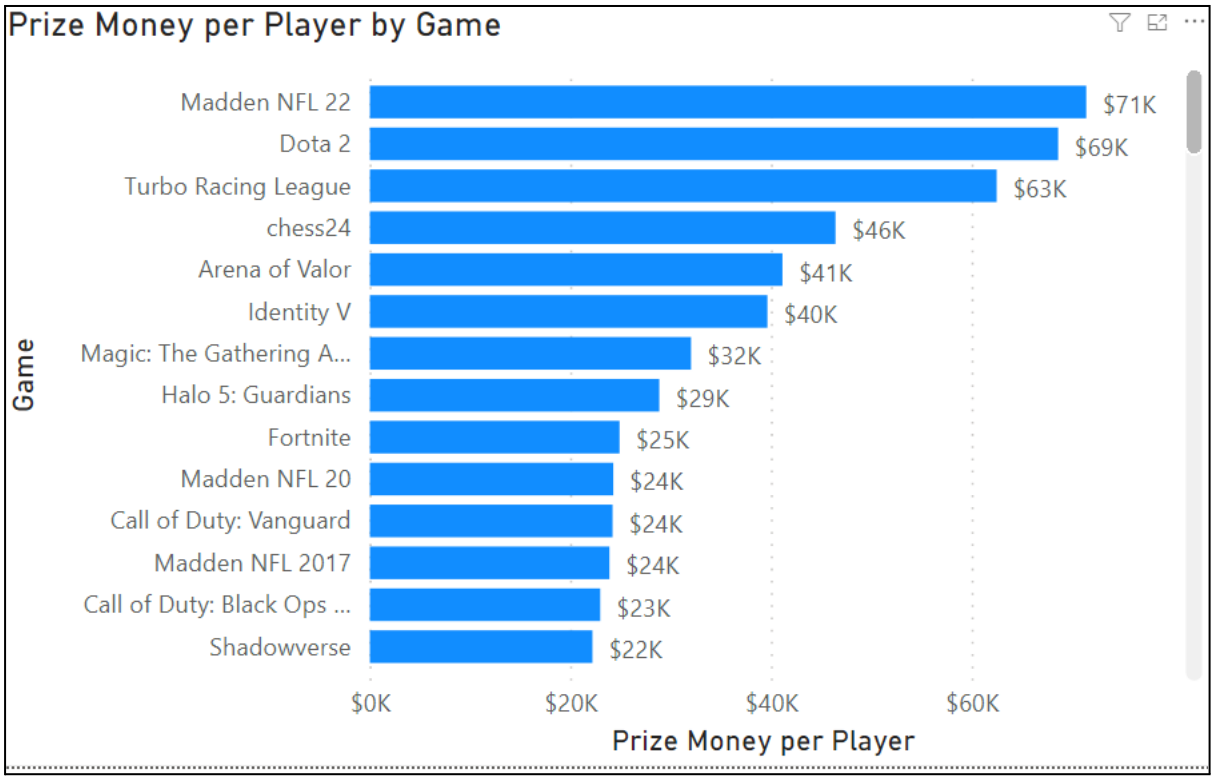
Studio as a data frame made it easy to clean the data by removing the rows and columns that were not necessary for our analysis and made it simple for us to verify the data was complete and functioning for our purposes. To help answer our follow up question, we needed year over year data. This required us to gather multiple datasets from sullygnome.com and import them into R Studio same as before, but then we needed to compile them into one dataframe for analysis. We were able to use the rbind() function within R Studio, which served our purpose perfectly. Lastly we wanted our data in CSV form to be used within Power BI for data visualisations, and we were able to use the write.csv() function to export our datasets.

Results

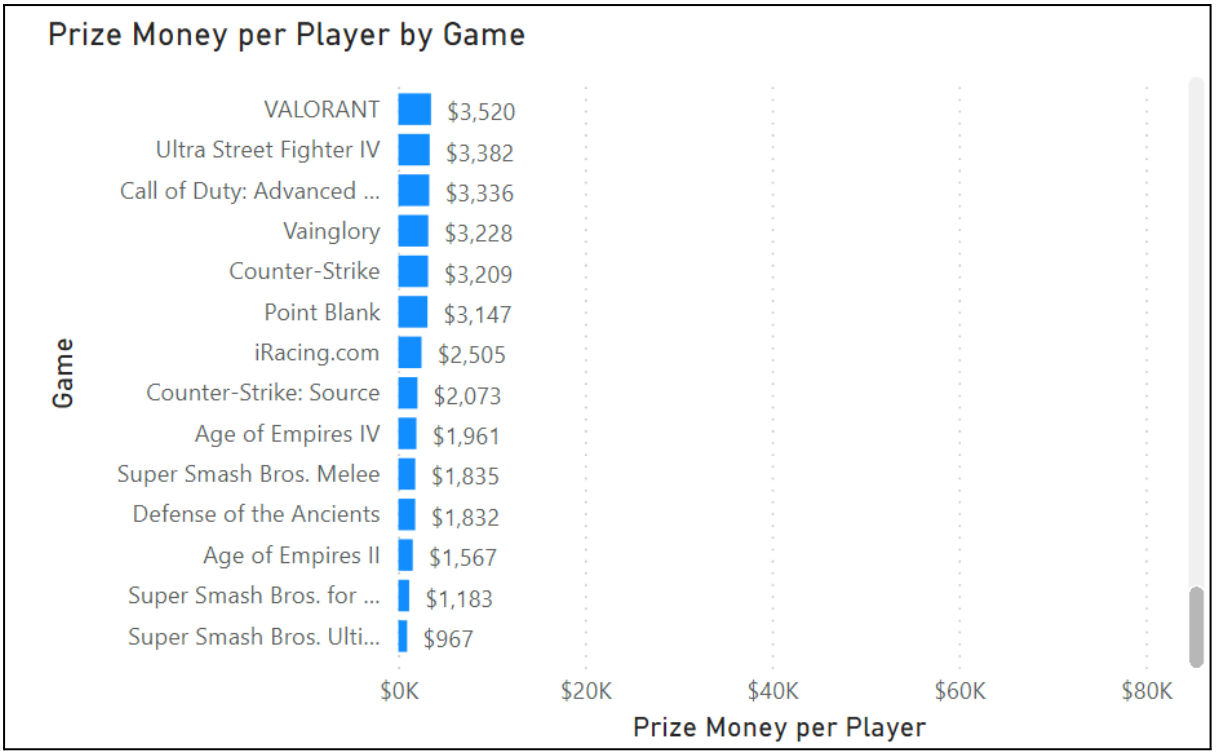
Our answer to question 1 of which game's eSports you should compete in professionally to make the most money is Dota 2, with a total prize pool over $300 million. The chart below shows the total prize awarded by game, along with the number of tournaments held, and the total number of players that compete in these tournaments. We see that Dota 2 has the highest prize money awarded, despite not having the most number of tournaments nor the highest number of players competing.



Examining the average prize money per player by game, the chart below shows that Madden NFL 22 has the highest average prize money per player by game. However, this can be considered outlier data as Madden NFL 22 has only had one tournament. Dota 2 is a close second, with the average player making around $69,000 worth of prize money.



Examining the other end, the games with the lowest average prize money per player are Super Smash Bros. Ultimate and Super Smash Bros. for Wii U, indicating that these games are highly competitive and that players looking to earn the most money should not compete in these games.



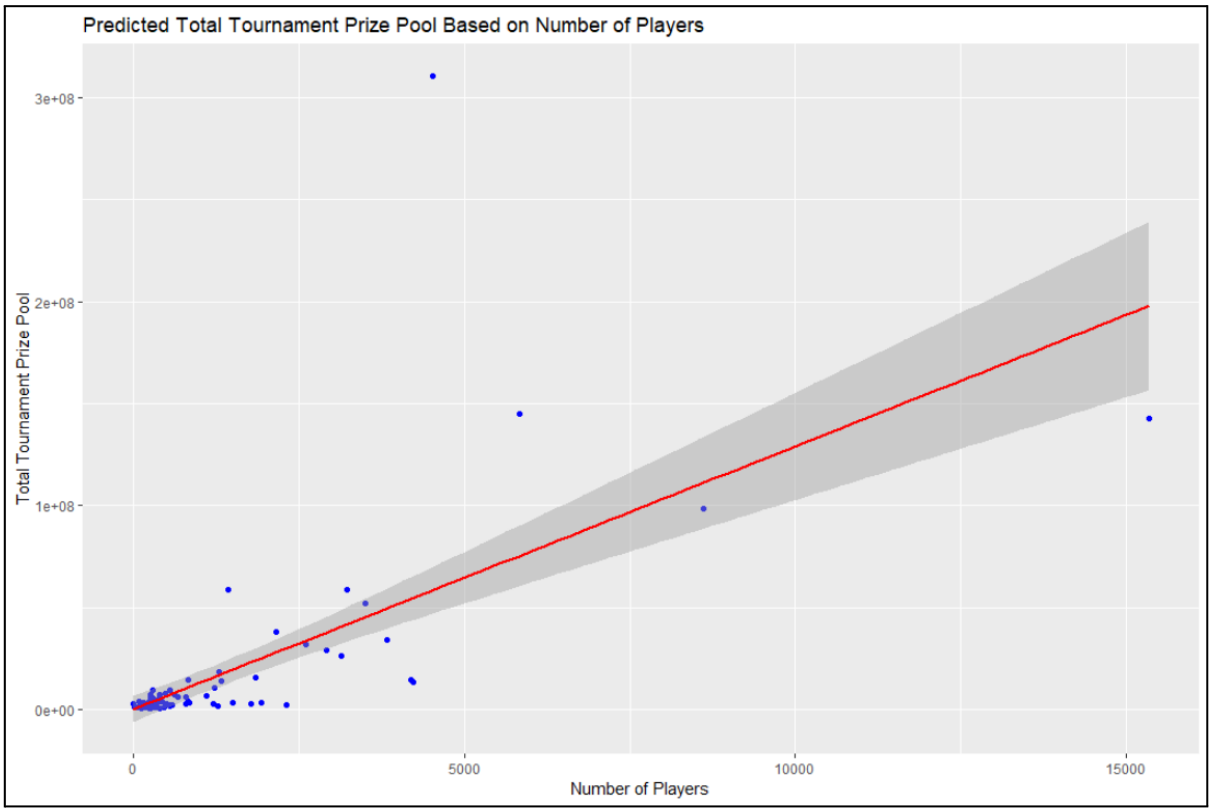
We also wanted to determine if there was a relationship between the number of players competing in a game and the total tournament prize pool. If this was the case, this would suggest that more popular games would also have more money to be divided among these players. To do this, we created a linear regression model to predict the total tournament prize pool for a game, using the number of players as a predictor. The summary of the regression is shown below

```
Call:
lm(formula = complete_info$Prize ~ complete_info$Players, data = complete_info)

Residuals:
      Min        1Q    Median        3Q       Max
-54925076  -4531031   -964699    572062 252262470

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           210242    3249395   0.065    0.949
complete_info$Players  12857       1441   8.920 2.67e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28710000 on 98 degrees of freedom
Multiple R-squared:  0.4481,    Adjusted R-squared:  0.4424
F-statistic: 79.56 on 1 and 98 DF,  p-value: 2.672e-14
```

The model creates a prediction following the formula:

$$Total\ prize\ pool\ =\ 177,513\ +\ 12,882\ *\ Number\ of\ players$$

In other words, for every 100 additional players competing in a game, we can expect the total tournament prize pool to increase by about $1.29 million. Plotting this regression line provides a regression graph as follows below:



We can see that generally, the points fit our regression. However, there are outliers that clearly do not fit our regression model, such as Dota 2 being the highest point. Our adjusted r-squared value of 0.448 reflects this, showing that there is at least some level of correlation, but not a particularly high one.

Our answer to question 2 of which team or organization has the highest amount of prize money won is Team Liquid, with nearly $43 million dollars won in prize money, in over 2300 tournaments played. The simplest way to interpret this is since Team Liquid has an incredible number of tournaments played, it would be reasonable to believe that they have

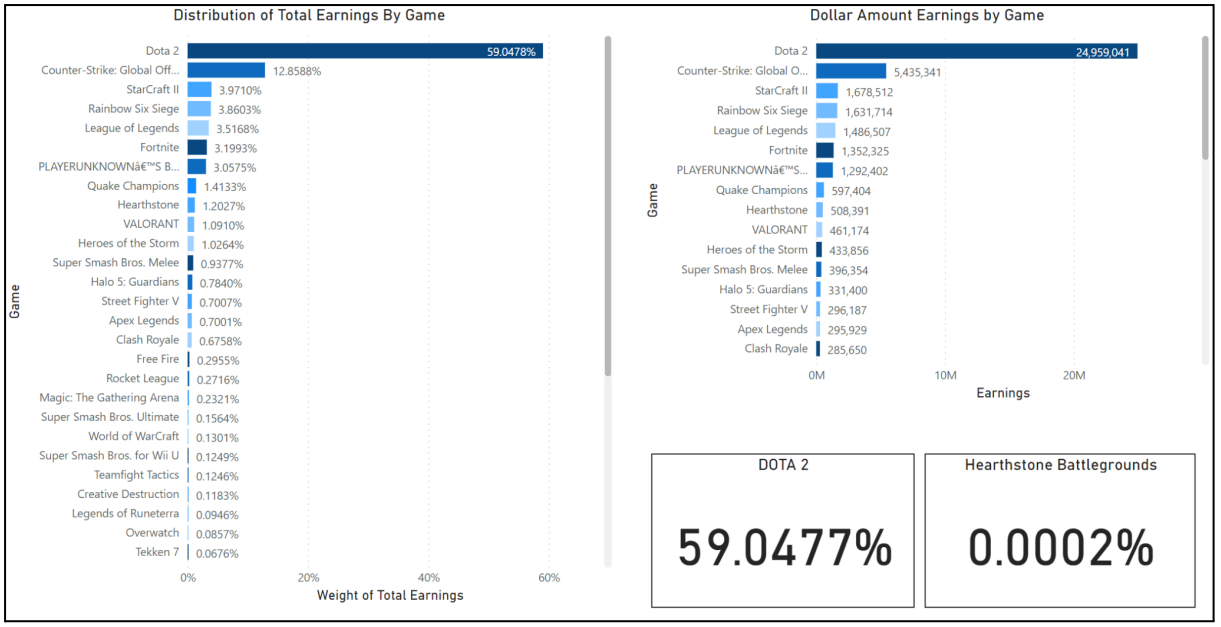
earned an incredible amount as well, assuming they are top performers at each event. With Team Liquid being founded in the year 2000, their long-time existence and experience with eSports contribute to their overall earnings.

```
# We know that the first row contains data for the highest earning team, so ...
# we can assign the total earnings for the highest earning team to a new variable.
highestEarningTeamMoney <- lookupHighestEarningTeamsDF$TotalUSDPrize[1]

# Then we can paste our answer displaying the team name and total earnings. As of ...
# December 1, 2022, the highest earning team globally is Team Liquid.
paste("The highest earning team is", lookupHighestEarningTeamsDF$TeamName[1],
      "with total USD earnings of:", prettyNum(highestEarningTeamMoney,
                                               big.mark = ",", scientific = FALSE))
```

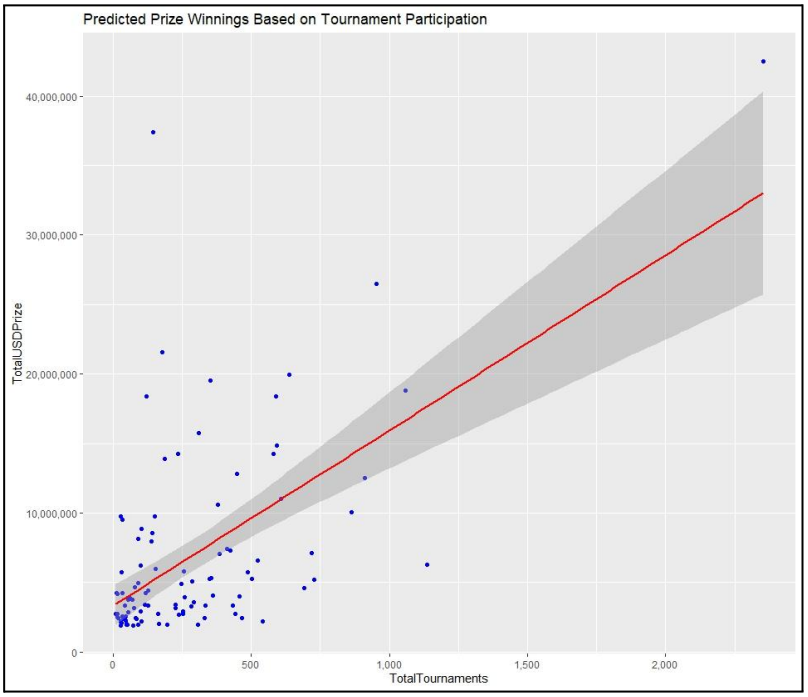
```
[1] "The highest earning team is Team Liquid with total USD earnings of: 42,491,961.57"
```

To answer our first follow-up question of what the percentage distribution of their total earnings is from all the games they participate in, we generated the following table and visualization:

	Var.1	Game	Total..Team.	Total..Overall.	X..of.Total	Team	percentPerGame
1	1	Dota 2	24959041.10	310549833.83	8.04	102-Team-Liquid	0.590477
2	2	Counter-Strike: Global Offensive	5435341.35	142703864.58	3.81	102-Team-Liquid	0.128588
3	3	StarCraft II	1678511.54	38293463.92	4.38	102-Team-Liquid	0.039710
4	4	Rainbow Six Siege	1631714.32	31901608.39	5.11	102-Team-Liquid	0.038603
5	5	League of Legends	1486507.25	98317652.24	1.51	102-Team-Liquid	0.035168
6	6	Fortnite	1352325.00	145185281.05	0.93	102-Team-Liquid	0.031993
7	7	PLAYERUNKNOWN'S BATTLEGROUNDS	1292402.30	52222180.38	2.47	102-Team-Liquid	0.030575
8	8	Quake Champions	597404.00	3279462.77	18.22	102-Team-Liquid	0.014133
9	9	Hearthstone	508390.89	28893344.27	1.76	102-Team-Liquid	0.012027
10	10	VALORANT	461173.74	14739648.21	3.13	102-Team-Liquid	0.010910
11	11	Heroes of the Storm	433856.39	18385772.65	2.36	102-Team-Liquid	0.010264
12	12	Super Smash Bros. Melee	396354.49	3542159.27	11.19	102-Team-Liquid	0.009377
13	13	Halo 5: Guardians	331400.00	7160480.97	4.63	102-Team-Liquid	0.007840
14	14	Street Fighter V	296187.23	2266798.43	13.07	102-Team-Liquid	0.007007
15	15	Apex Legends	295929.11	15479986.99	1.91	102-Team-Liquid	0.007001



To interpret this, we can see that Team Liquid's highest-earning game is Dota 2, comprising nearly sixty percent of their earnings, totaling roughly $25 million USD. In contrast, Team Liquid's lowest-earning game is Hearthstone Battlegrounds, which fails to appear on the graphs. It comprises 0.0002% of Team Liquid's earnings, totaling roughly $100 USD. We can also see that there is a major drop off between Dota 2 and the next most earning game, Counter-Strike: Global Offensive at nearly thirteen percent, or roughly $5.5 million USD. To accurately display our regression, we needed to generate the appropriate graph. The following is the result of our regression:



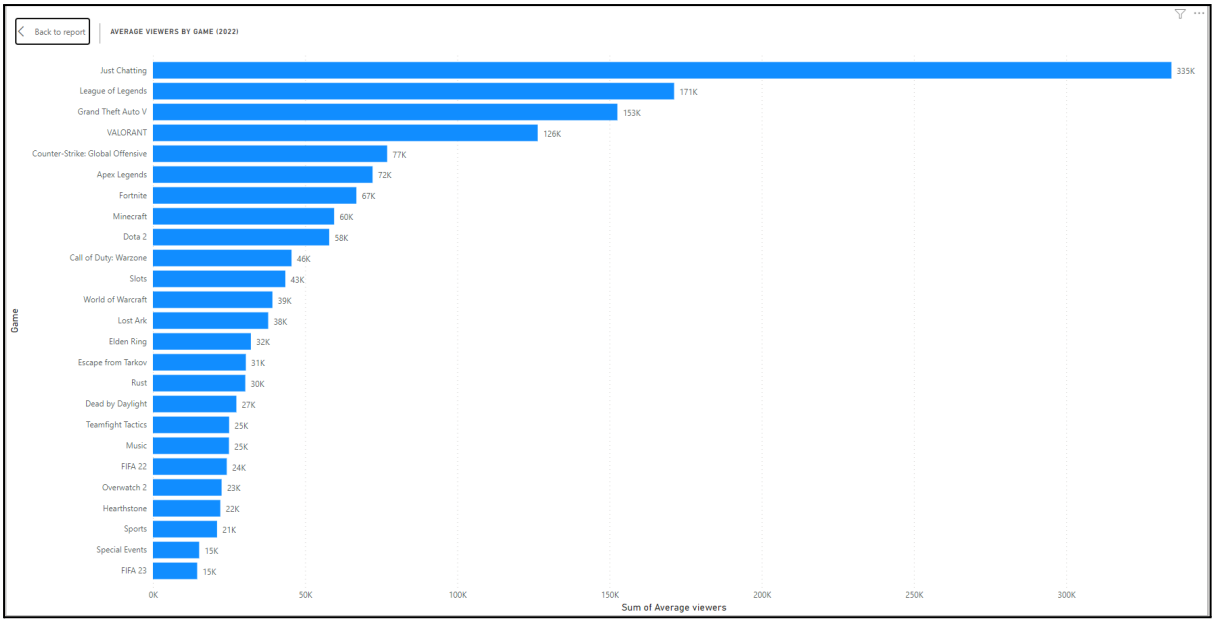
To interpret these results, we concluded that the model creates a prediction following the formula:

$$log(Total\ USD\ Prize) = 13.8297 + (0.3140 * log(Total\ Number\ of\ Tournaments))$$

What this means is for every one percent increase in the total number of tournaments, we can expect to see a 0.3140% increase in the total prize winnings. When inspecting these results, it falls in line with our expectations. Simply put, the teams that don't participate in as many tournaments, do not have as much prize money won as other teams. While our p-value for this predictor was acceptable, we found that our adjusted r-squared was less than desirable. We attribute this fairly low value to our lack of data and skewed results.

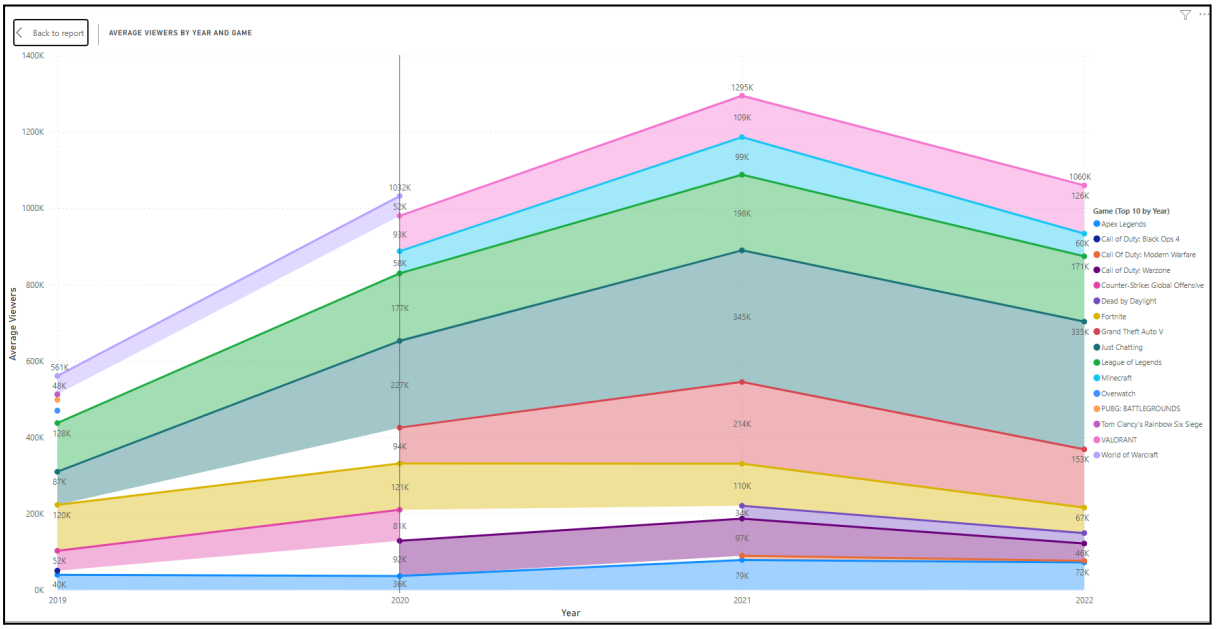
```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     13.82963    0.28950  47.771  < 2e-16 ***
log(lookupHighestEarningTeamsDF$TotalTournaments)  0.31404    0.05582   5.626 1.75e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6603 on 98 degrees of freedom
Multiple R-squared:  0.2441,    Adjusted R-squared:  0.2364
F-statistic: 31.65 on 1 and 98 DF,  p-value: 1.749e-07
```

Answering question three was best done by creating a horizontal bar graph (seen below) which showed the top 25 most viewed games on Twitch by average viewership. The first thing that came up as a result of our graph was interesting in that the top viewed topic on twitch was not a game at all, but was Just Chatting, a form of streaming that prioritises communication and conversation over actual gaming. Looking at the graph, it can be seen that this topic had nearly double the average viewers than the next highest, which was an actual game. While we found this to be interesting, the question we were seeking to answer pertained to gaming categories, so we looked beyond the top most viewed topic. The top three highest games found themselves in the second through fourth highest in terms of viewership. They went in order of: League of Legends at 171 thousand, Grand Theft Auto: V at 153 thousand, and Valorant at 126 thousand viewers. After these top three there was another steep decline in terms of viewership, which was for Counter-Strike: Global Offensive at 77 thousand which rounded out the top 5 most viewed topics on twitch in 2022.



Our follow up to question three was answered by creating a stacked area chart (seen below) of year over year average viewers by the top ten most viewed games in said year from 2019 to the current year. Based on our findings, we were able to conclude that, yes, COVID-19 did indeed have a significant impact for streamers by increasing their viewers. Looking at the first two years on the graph, it can be seen there was a significant increase in viewers from 2019-2020 as the average viewers among the top ten most viewed games nearly doubled from 561 thousand to 1.03 million. COVID-19 lockdowns began in March 2020 and we believe this to be a major factor in this increase. While 2020 to 2021 again saw an increase in viewers, it was not nearly as large, only seeing about a 25% increase over the previous year. This tells us that the impact of COVID-19 was significantly positive for Twitch streamers as people were not able to leave their houses during lockdown and found themselves spending more time online watching favourite creators on Twitch. Another finding we had pertained to the top ten games over the four year span. There were only 16 unique games among the top 10 games, which shows us that while there is some variability year over year as new games come out and different games are popular, the top ten games do stay relatively stable. Lastly, while the 2022 data is not yet complete and does not have December stats as the year has not yet finished, our thoughts and analysis is that there may possibly be a decline in average viewership among the top 10 games on Twitch compared to

last year, which once again proves our theory that COVID-19 had a major impact on Twitch viewership now that lockdowns have ended and people are returning to work and having less time to be online.



Discussion

Overall for question 1, the main takeaway is that games with more players competing in them generally have more prize money. There are however exceptions to this, such as Dota 2, which has a large prize pool for the number of players it has, and also Super Smash Bros. games, which have the lowest prize money per number of players. By competing in games that are less competitive, or have larger prize pools given the number of players that compete in these games, one can look to maximise their earnings. However, even towards the top, the average lifetime tournament earnings of $70,000 is hardly enough to support anyone, suggesting that while esports is a growing industry, it is still highly unlikely that one will be able to live off tournament earnings alone.

Overall for question 2, the main takeaway is that the more you participate, theoretically the more prize money you will win. Drawing from what we see in Team Liquid, there are other factors that contribute as well such as their time spent existing as an organization, as they were originally formed in 2000. Coupled with the fact that they have

over 2300 tournaments played also puts them as one of the most competitively experienced organizations in the industry. Team Liquid is the true embodiment of “play more, win more”, and is something that all prospective teams and players should adopt (assuming they perform competitively enough to challenge other teams and players).

However, from our results, we can also see that there are teams that perform exceptionally well, and well enough to break out of our predictions. Teams like OG and Team Spirit have all earned a comparable amount (over 50%) of Team Liquid’s earnings, despite competing in less than one-**TENTH** of the number of tournaments that Team Liquid has. Furthermore, both of these organizations were founded recently in 2015! In the span of seven years, both of these teams have accomplished half of what Team Liquid has accomplished in their 22 years of existence. What this means is that for prospective teams and players who feel incredibly out-classed by the likes of Team Liquid, Evil Geniuses, and NaVi; there is no need to fear! With hard work, dedication, and an eye for talent, newcomer teams and players alike have the chance to be extraordinary and compete at the same level as other seasoned teams. For established teams like SK Gaming, CLG, and Mouz, who are all similar in age to Team Liquid but have won only a fraction in comparison, this could be a sign to start mixing things up. Change which games they invest in, their rosters, or even take a page out of OG’s book, as you could even learn a thing or two from the new kids on the block.

For question three, the main takeaway we had is that, while Twitch is commonly thought of by most as a streaming platform for gamers to share live gameplay and gaming content, and yet the top viewed topic on Twitch is a category that does not focus on gaming.This could be an interesting discovery to streamers. While you might think that you can just categorise your stream as Just Chatting and play the game you would like, current Twitch community guidelines state that you are expected to accurately label your content to the best of your ability (Twitch, n.d.). We see some loopholes to this for streamers, like categorising their game as Just Chatting, and making the gameplay not the main focus of the stream. We believe this could be useful to smaller streamers wishing to grow their

channel, especially if the game they choose to stream is not within the top ten in terms of viewership or has less viewers overall. While this could oversaturate the Just Chatting topic, it could be a useful workaround until the channel grows to the point they don't need the Just Chatting topic to gain viewership. We believe this could be a good option for the time being so long as the guidelines do not change.

References

Esportsearnings. (n.d.). *Home Page*. Retrieved December 1, 2022 from https://www.esportsearnings.com/

Sullygnome. (n.d.). *Home Page*. Retrieved December 3, 2022 from https://sullygnome.com/

Twitch. (n.d.). *Community Guidelines: Introduction to Safety on Twitch*. Retrieved December 3, 2022 from https://safety.twitch.tv/s/article/Community-Guidelines?language=en_US