Analysis of Trends in E-Sports

Jacob Collins, Thomas Devine, Lai Le, Diana Tyler
CSU, Chico: CSCI 385, Group 9

Introduction

eSports is a form of high level gaming that involves competing for prize money or notoriety. eSports are played in the format of team vs team, or solo player vs solo player. Many of the most popular games can have tournaments with prize pools up to a million dollars. eSports creates a new potential market for game developers to target, with developers being one of the main groups to consistently profit from a thriving eSports scene, as well as several teams having created valuable companies off the back of eSports success. Given all of this information, our team is taking on the problem of determining patterns within the eSports and overall video game infrastructure that can work to benefit groups such as players, developers, and teams as much as possible. eSports is a field where many people find community and enjoyment as well as being able to make careers, and this analysis could help improve the lives of all those people. We have aggregated data from several sources such as EsportEarnings.com and sullygnome.com and performed analysis on this data to try and address our fundamental data science question: What does general eSports data show us about how teams, players, and developers can invest their time and resources to find the best return on their investments?

Related Work

Previous work related to this area includes participating in a proponent/skeptic discussion of papers related to our subject. The team had to select a piece of data science work related to E-sports and highlight its good and bad qualities. We chose the paper “Knowledge management in the esports industry: Sustainability, continuity, and achievement of competitive results,”[4] which goes heavily into the details of esports and how the coronavirus changed it. Having to study the paper for the discussion, we studied the trends. For example, E-sports should continue to experience tremendous growth even after the pandemic. By studying the paper, the team further grew their knowledge on the strengths, weaknesses, opportunities, and threats on E-sports. After the discussion, the team is able to read and understand pieces of data science work while maintaining a good amount of skepticism.

Methods

EDA

We clarified the meaning of all variables in our datasets before analyzing a summary of these variables to greater depth. The summaries for TeamId, TotalTournaments, and TotalUSDPrize were a particular source of some confusion, as they were listed in different instances by month and by game. We mutated the dataset with some filters and matching to get more broadly-scoped data that could be used to visualize the trends in E-Sports as a whole, most notably by converting the monthly increments to yearly.

We removed the 2023 data from our dataset as it was incomplete and would produce misleading visualizations if included presently.

We also noticed that TeamId had many counts, and came to the realization that many E-Sports teams participate in tournaments for many different games, often more than 5. There are also two main groups of TeamId, but we decided that this was not worth further consideration as there was no improper overlap in names between the two groups (0-300 and 2300-2500), and we would gain no benefit from reorganizing the TeamIds.

We plotted distributions of how many tournaments each individual team participated in, but decided that it was too fine-grain for a quality presentation, and so opted to instead group these tournament counts by the top 10 games and create a box plot.

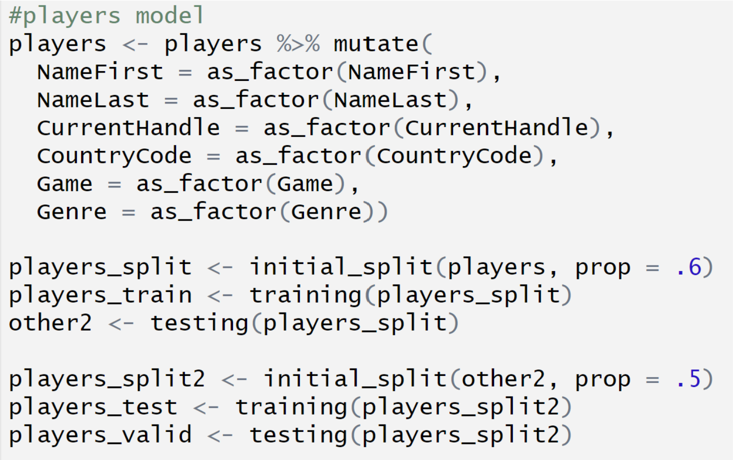
One of the more informative figures came from plotting the number of tournaments that teams in a certain game participated in for that game, while the total earnings in millions USD was presented behind. It can be a little confusing at first, but viewing the differences in these trends can give valuable insight to potential investors on the earning efficiency of different games.

Mutating the historical dataset from a monthly schedule to a yearly schedule was a good learning experience, as it involved merging several columns from our general E-Sports data set, parsing the date for each tournament event, converting it to the beginning of that year, converting that back to a date, and adding the new date column, all in a way that does not disrupt the rest of our data. This was a successful process and was very helpful moving forward.

Lastly, we looked at the yearly total tournament earnings of games in each genre to see what the financial environment is like from a broad perspective. This was done by setting up a loop to go through all unique values of genre, and generating a plot from the data filtered to only match one specific genre at a time.

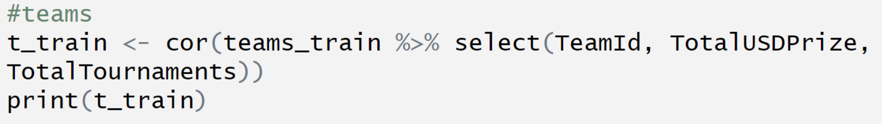
Model

We started off by doing a little bit of cleaning before setting up our models. The clean up was made to remove garbage values that did not make sense and were likely to skew our data. An example of this can be seen in our general data which holds 8 variables in which two of them are Total Players and TotalTournaments. There are data entries like “Dead by Daylight Mobile” which has 0 players yet has 1 tournament. Another data entry is “Descent 3” which has 1 player and 1 tournament. We just did a little filtering in which we started to set up our 6 datasets which are: teams, players, locations, general, historical, and twitch-games.

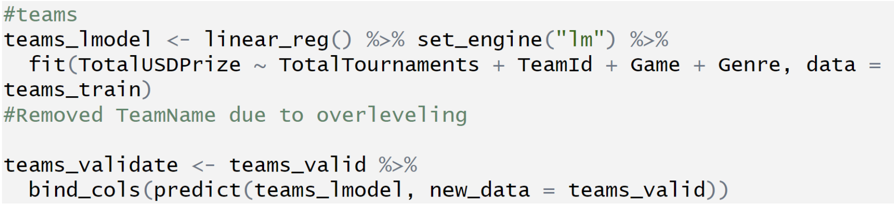


```
#players model
players <- players %>% mutate(
  NameFirst = as_factor(NameFirst),
  NameLast = as_factor(NameLast),
  CurrentHandle = as_factor(CurrentHandle),
  CountryCode = as_factor(CountryCode),
  Game = as_factor(Game),
  Genre = as_factor(Genre))

players_split <- initial_split(players, prop = .6)
players_train <- training(players_split)
other2 <- testing(players_split)

players_split2 <- initial_split(other2, prop = .5)
players_test <- training(players_split2)
players_valid <- testing(players_split2)
```

We started off by separating our data into 3 sections: training 60%, testing 20%, and validating 20%. The method used for this was with initial_split in which it first splits the data to 60-40 then the final split which is 50-50 on that remaining 40. This allows us to set up our models into three which can be used to validate our results. During this split, we ran into some issue as the majority of our variables for these datasets were either characters or dates. We had to convert them into factors so they can be properly trained and for dates we use as.numeric and as.character to achieve the same results.



```
#teams
t_train <- cor(teams_train %>% select(TeamId, TotalUSDPrize,
TotalTournaments))
print(t_train)
```



```
#teams
teams_lmodel <- linear_reg() %>% set_engine("lm") %>%
  fit(TotalUSDPrize ~ TotalTournaments + TeamId + Game + Genre, data =
teams_train)
#Removed TeamName due to overleveling

teams_validate <- teams_valid %>%
  bind_cols(predict(teams_lmodel, new_data = teams_valid))
```

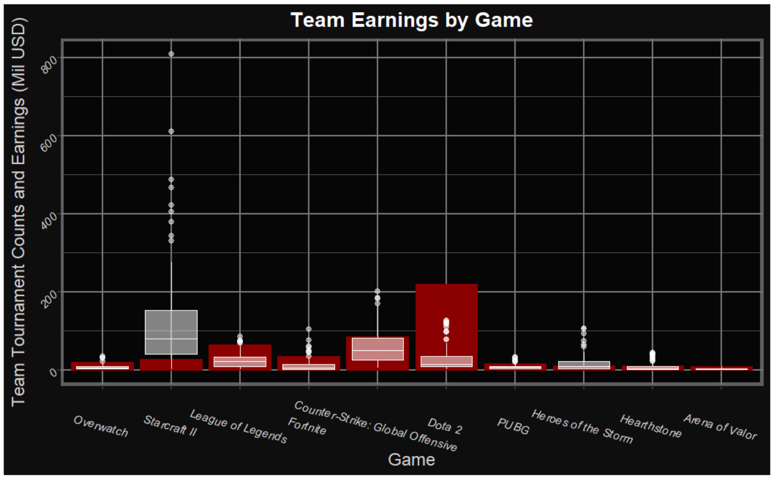
Now that we have our models set, we tried to find the correlation between the numeric attributes within the datasets. This was one of the feedback we got from the peer review. To show the correlation of the variables for the user to get a better understanding of the dataset. Even though about half of our variables are now factors, this allows us to see which variable to choose when deciding on which variable will be used for the linear regression. Doing the linear regression was straightforward, but it was where we had the majority of our issues. These are likely the reason why our data and graphs were off due to having half of the attributes as factors. Since the attributes were factors, it was constantly causing our linear model to have overleveling issues. Which means that there are certain values that are not present in either of the three models. For example with our locations model, a common issue was that there were certain states that were not available in all of the models. Like the training model did not have ‘AZ’ or ‘CA’, but the validating model did. We tried to fix it by filtering out those overleveled values, but we could not get it to work properly. A solution that we had was to manually insert all the overleveled values into the models, but that would be too time consuming especially with all the workload our team had. So, we decided to drop the variables which gave us our wonky data. Like previously mentioned, a lot of our data had characters for variables which meant that we dropped a lot of our variables. However, we did get some nice looking data and their graphs from certain datasets such as twitch-games, general, and player model. The worst looking one was the location model which makes sense with what was brought up earlier.

```
rsq <- yardstick::rsq(teams_validate, truth = TotalUSDPrize, estimate =
.pred)
rmse <- yardstick::rmse(teams_validate, truth = TotalUSDPrize, estimate =
.pred)
mae <- yardstick::mae(teams_validate, truth = TotalUSDPrize, estimate =
.pred)

results <- workflow() %>%
  add_model(linear_reg()) %>%
  add_formula(TotalUSDPrize ~ TotalTournaments + TeamId + Game + Genre) %
>%
  fit_resamples(vfold_cv(teams_train, v=5),
              metric_set(rsq, rmse, mae))

collect_metrics(results)
```

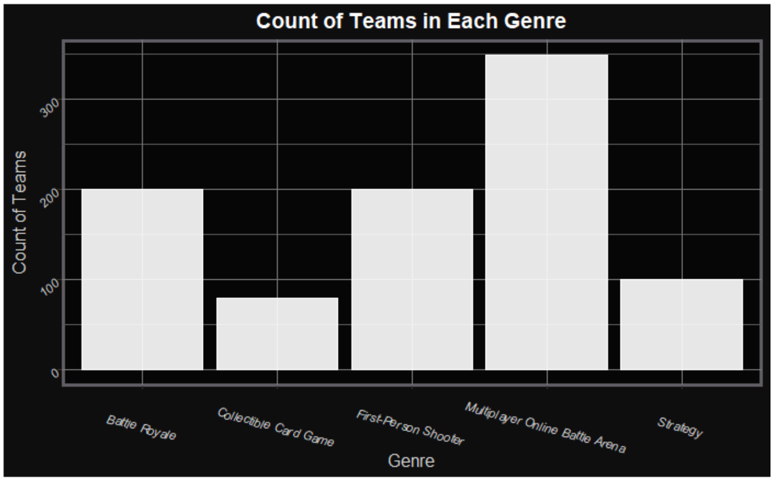
Finally, we then found the rsq, rmse, and mae of each training/validating models for each dataset. We then used that to solve the k-fold cross-validation for each model. After that, we then compare the results to see the accuracy of our models and tests which we were pretty happy with for 4 out of our 6 datasets. There were different approaches that we could have done and we would like to if given enough time. The models could have been and can be improved further if we would like to continue on with this project.

Results/Discussion

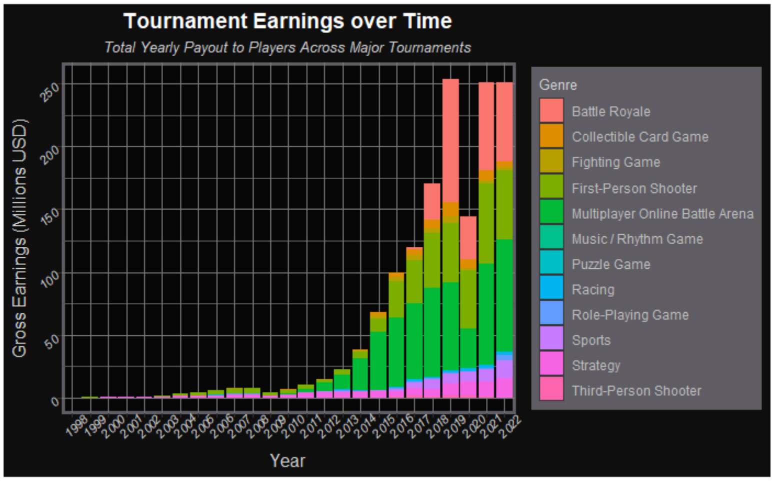
EDA



There are some interesting discrepancies between how active a game's tournament scene is, compared to the total tournament earnings rewarded for that game. For example, Starcraft II has an astonishingly active tournament scene, but these teams receive relatively little earnings compared to other games. In contrast, Dota 2 has a fairly average number of tournaments per team, but the payout for these tournaments is far higher than most of the other games on display. It is worth keeping this relationship in mind, as potential investors would likely wish to maximize the earnings per game, rather than just the total potential earnings.

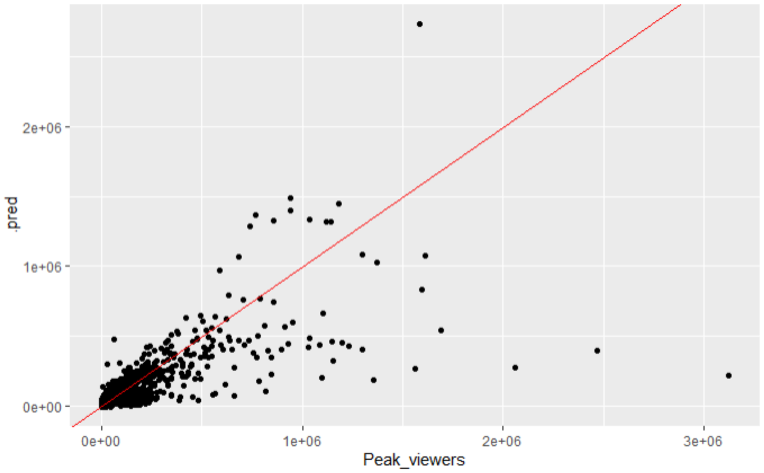


A game with more active and unique teams may be easier to get into and start participating in tournaments for, but may perhaps also be more competitive. Understanding the landscape of a particular genre in E-Sports is a crucial step towards differentiating between what may be a valuable or risky team to invest in.

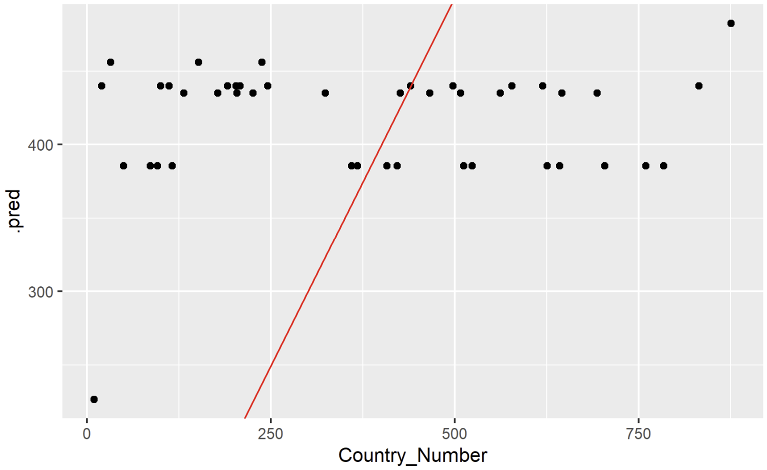


A significant point of concern regarding E-Sports in recent years is the effect of the pandemic on this industry. We can clearly see there was a nearly 50% decrease in total tournament earnings in 2020 as E-Sports tournaments struggled to shift towards a fully-online format, but it should be easy to tell that the industry has since recovered, and E-Sports teams are receiving nearly the same total payout as they had prior to the pandemic. It is worth noting that this year there have been multiple record-breaking tournament payouts, and it is expected that 2023 will be a new high-score for E-Sports tournament earnings.

Model

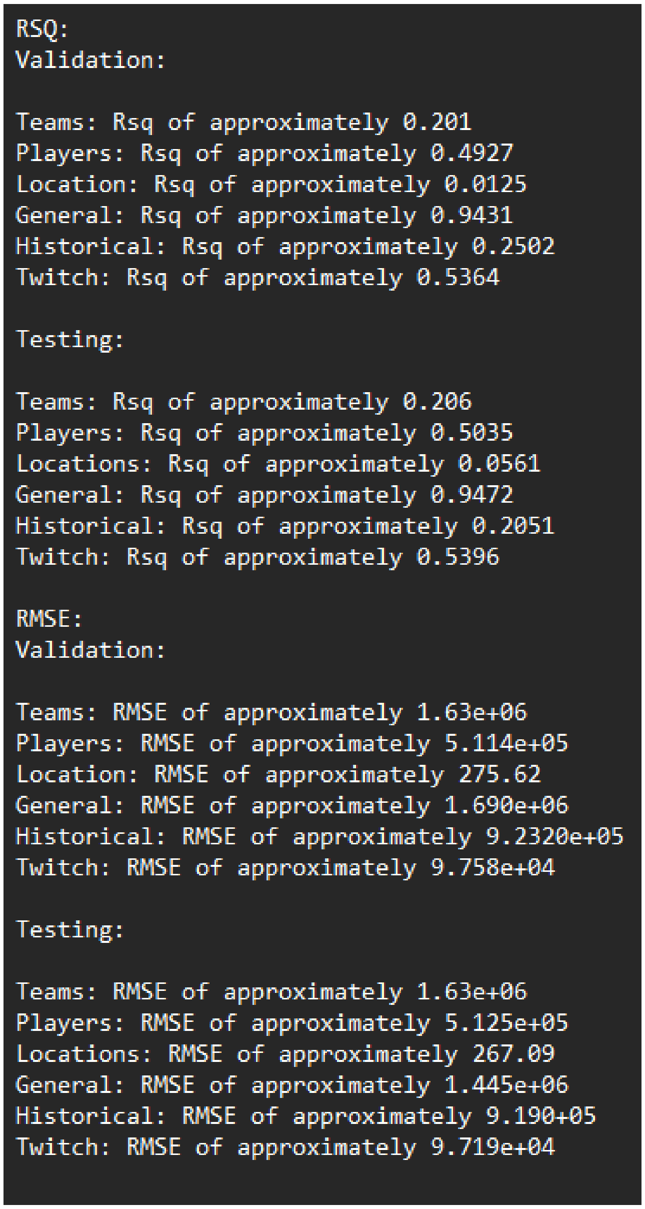


This is our twitch_game model which shows a strong prediction between the chosen variable and the rest of them under .pred. This was one of the better models which showed that the majority of the variables within this model were integers which helped with our predictions. The graph illustrates that this model's prediction is very close to the actual values. As for the dots that are not close to where the dots congregate or where the line is, those are the outliers which causes the predictions to deviate from the actual models. These outliers can create inaccuracies within the model, but they are important as it shows which games are earning the most money within our datasets.



This is our location model which is also our worst model. Due to the fact that the majority of the variables were characters, we had to drop several of our variables for it to work. We could have solved it another way by probably joining all our datasets together and filtering out the overleveled value before working on the model, but we were limited on time. As one can see, the model shows a poor prediction as our plots are everywhere and not close to our red diagonal line. This model is inaccurate and should be avoided unless we can figure out a different model to test it with instead of a linear regression.

For the results, the validating and testing models results were really close to one another. This was solved using K-fold cross-validation. These results can be interpreted differently based on size of the data, but generally the smaller the RMSE values indicate that the model’s predictions are closer to the actual values and RSQ is the other way around where the higher the value, the more accurate it is which indicates that this is a better fit of the model. Below is an image with all our values based on the models. For RSQ, our dataset is a good indicator of the variance with the variables and locations is the worst indicator for this. For RMSE, our models have large rmse values which suggests that our predictions are going to have more errors compared to the actual values. However, they are pretty consistent between the two models for each RSQ or RMSE.



```
RSQ:
Validation:

Teams: Rsq of approximately 0.201
Players: Rsq of approximately 0.4927
Location: Rsq of approximately 0.0125
General: Rsq of approximately 0.9431
Historical: Rsq of approximately 0.2502
Twitch: Rsq of approximately 0.5364

Testing:

Teams: Rsq of approximately 0.206
Players: Rsq of approximately 0.5035
Locations: Rsq of approximately 0.0561
General: Rsq of approximately 0.9472
Historical: Rsq of approximately 0.2051
Twitch: Rsq of approximately 0.5396

RMSE:
Validation:

Teams: RMSE of approximately 1.63e+06
Players: RMSE of approximately 5.114e+05
Location: RMSE of approximately 275.62
General: RMSE of approximately 1.690e+06
Historical: RMSE of approximately 9.2320e+05
Twitch: RMSE of approximately 9.758e+04

Testing:

Teams: RMSE of approximately 1.63e+06
Players: RMSE of approximately 5.125e+05
Locations: RMSE of approximately 267.09
General: RMSE of approximately 1.445e+06
Historical: RMSE of approximately 9.190+05
Twitch: RMSE of approximately 9.719e+04
```

Response to Feedback

- We received multiple comments about how our models were unclear. We struggled with balancing clarity, and aesthetic through many of our figures.
- Our EDA was much different than our final presentation because of the focus of each. With our EDA the group decided to work on discovering current trends in the industry. This is not particularly relevant because of the trends of genres, games, etc. Due to our feedback, we decided to pivot our project to be about future trends in the Esports industry.
- This was invaluable feedback that worked in our favor as it became a more interesting project and we were able to predict the trends in the industry with a moderate amount of success.
- Our final piece of feedback was the clarity of the focus of our project. We decided to go over what our project was, add relevant examples of our data, along with our graphs.

Future Work

- It could be nice to plot some tournament efficiency metric rather than having team tournament counts overlaid with total tournament earnings per game.
- Taking ad revenue into consideration would be nice. Potential investors may want to know how marketable teams for a certain game or genre may be, as an additional revenue stream, or as a reason to sponsor them in the first place.
- Average team revenue would also be an interesting statistic to see, particularly filtered by game, genre, or country.
- Some kind of residual for the histograms showing total yearly tournament earnings per genre would be interesting. Perhaps a difference from the average of all yearly earnings could allow us to more closely see what peculiarities there are specific to a particular genre.

Conclusion

Through our explanation and analysis of this data, we found multiple factors that could be interesting or important for eSports teams, players, and developers to find success. It is clearly very important to understand the landscape of the competitive scene for the type of game being invested in, some genres like MOBAs or Battle Royales boast high prize pools and many players but also a high amount of competition, meaning standing out in these crowded markets may be difficult and likely requires a significant amount of resources and investment. On the other hand, genres like fighting or racing games may take a smaller slice of the pie in terms of earnings, but these communities are smaller and have less titles to choose from so they may be more willing to give a new game from a smaller developer or indie developer a chance. The Covid-19 pandemic did have a negative effect on eSports, but from our metrics things seem to be stabilizing or even surpassing previous years of money made and eyes on the eSports scene. Though we had some struggles with fitting our various sets of data to a linear model, we ultimately were able to reach meaningful and interesting results.

Source code for our data analysis can be found on GitHub.

Contributions

Lai Le

- EDA country/genres
- Presentation, Report
- Tidying data

Diana Tyler

- Contributed to data science questions and exploratory analysis
- model validation
- structure for presentation/reports

Jacob Collins

- EDA Earnings by Game
- EDA Time series’ by Genre
- EDA Earnings by Genre
- Linked datasets
- Converted monthly to yearly times.

Jesus Alvarado

- Training Models
- Filtering Datasets
- EDA results

Thomas Devine

- EDA formatting
- EDA Graphs
- Presentation Formatting
- Data Interpretation

Citations

[1] Daoud, J. (2020, December 17). Esports Earnings. Kaggle. https://www.kaggle.com/datasets/jackdaoud/esports-earnings-for-players-teams-by-game

[2] Esports Earnings. (2023). Prize money / results / history / statistics. Esports Earnings. https://www.esportsearnings.com/

[3] Kirsch, R. (2023, October 10). Esports earnings 1998 - 2023. Kaggle. https://www.kaggle.com/datasets/rankirsh/esports-earnings

[4] Saiz-Alvarez, J. M., Palma-Ruiz, J. M., Valles-Baca, H. G., &amp; Fierro-Ramírez, L. A. (2021, September 30). Knowledge management in the esports industry: Sustainability, continuity, and achievement of competitive results. MDPI. https://www.mdpi.com/2071-1050/13/19/10890