Data Selection

Our dataset is Steam Community Market Data from CSGO/CS2. We chose this dataset because we were originally interested in predicting market trends and seeing how different models perform with this task.

However, after our discussion with Pierce, we decided to switch our topic slightly. Now we are interested in seeing if we can predict winning teams, or top players, from CSGO/CS2 tournaments by viewing market data during a window of time before the tournamenet up until the last day of the tournament.

This dataset was obtained by sending HTTP requests to the following URL: `https://steamcommunity.com/market/pricehistory/?appid=730&market_hash_name={ITEM}` where `ITEM` is the name of a CSGO/CS2 item that has been url encoded.

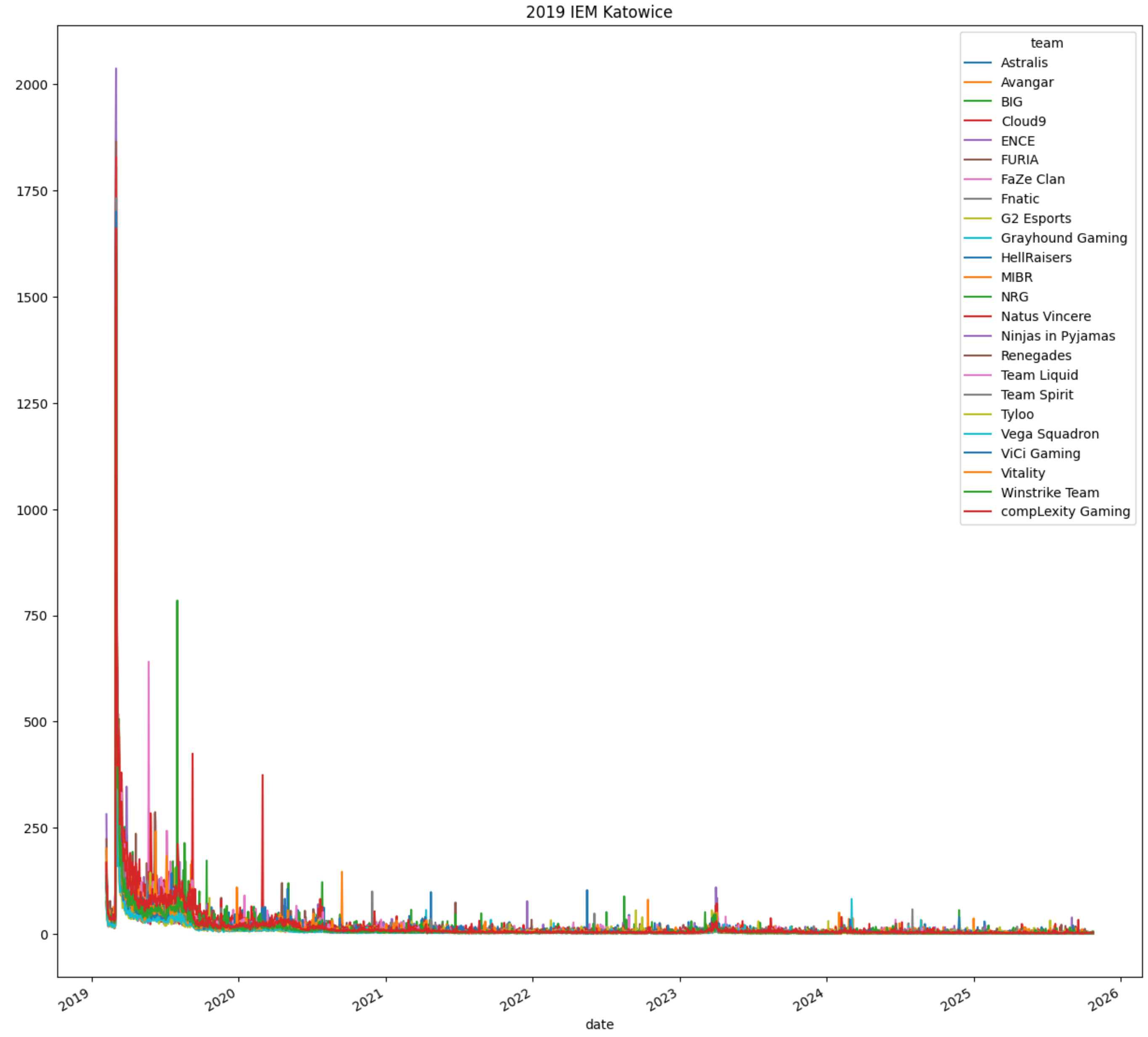
This data was then combined with results from an unofficial JSON API: `https://raw.githubusercontent.com/ByMykel/CSGO-API/main/public/api/el/`, which we found at the following link with instructions `https://bymykel.com/CSGO-API/#introduction`

Lastly, we gather CSGO/CS2 tournament data by using **Selenium** to scrape the `https://hltv.org` page for various tournaments. These tournaments were

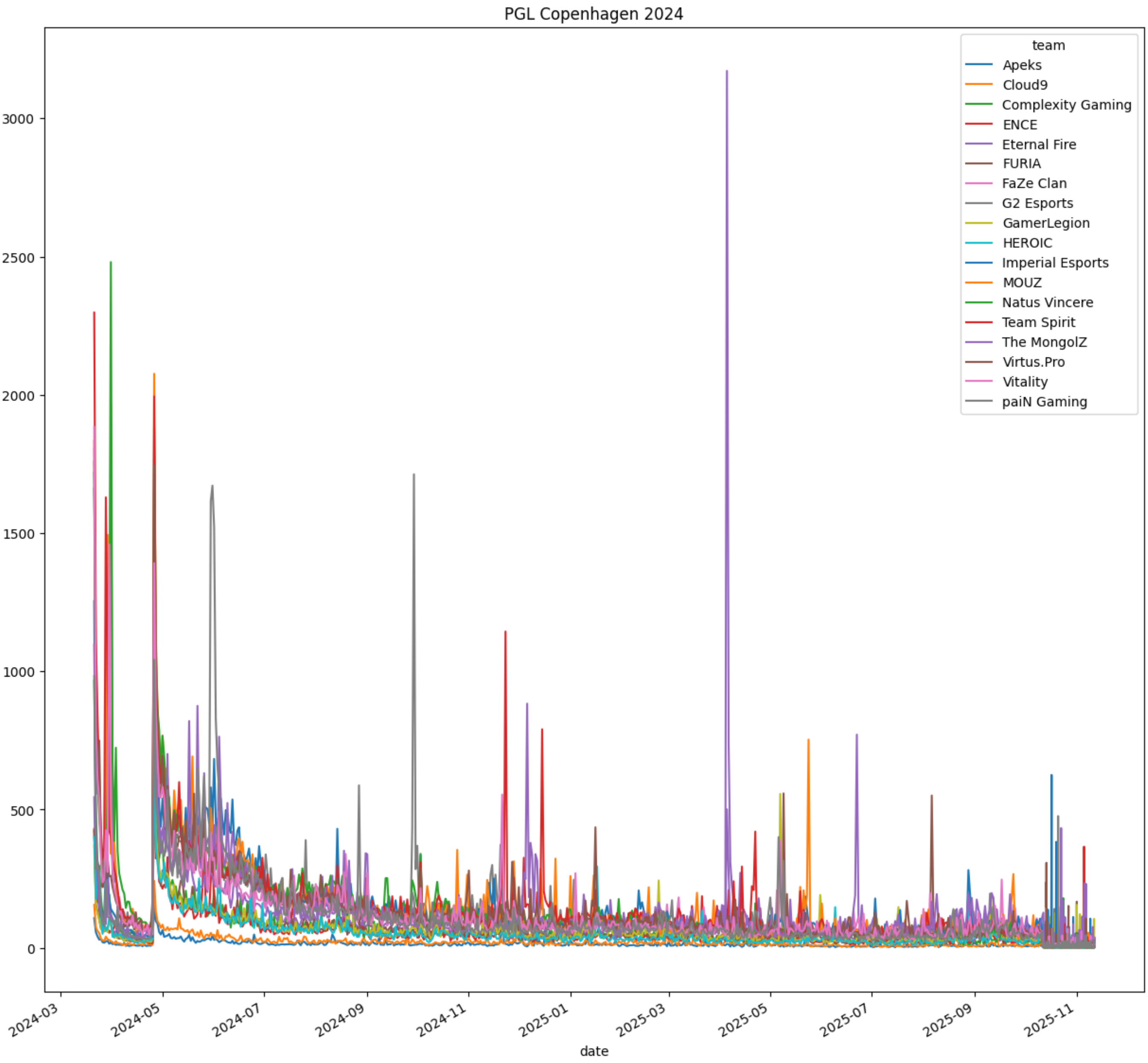
The official Steam Community Market provides the median price and volume sold in the following way:

- Hourly for data within the last month of the request
- Daily for data that is over a month old

Exploratory Data Analysis (EDA)xw



Here's an example of the distribution of volume of stickers for a tournament period, `IEM Katowice 2019`. The winners of this tournament was team `Astralis`, with second place being team `ENCE`. This graph was a little more questionable than the previous. There is a very defined peak on the left-hand side of the graph that, yet the tallest line corresponds to `ENCE`. We need to look further into this, but our hypothesis is that as time approaches the final matches, sticker sales correspond to community predictions.



Lastly, here's an example of a distribution where predictions become very difficult. This was `PGL Copenhagen 2024`, where the winning team was `Natus Vincere`. We can see on the far left that there are two primary peaks that roughtly correspond to the top placing teams, including Natus, but after the tournament ends we still see these peaks occurring that don't correspond to any of htese top teams. Especially in around April of 2025, we see a large spike in sales for a team that did not even end up in the top 5. This tells us that when doing predictions, we will need to narrow our scope of time.

Modeling and Feature Development

The goal for our models is to predict the Top 3 teams of a given tournament by using the steam market sale data.

The items we are using are all `Sticker` items that are associated with a specific team that was in the given tournament.

Evaluations are performed with `Precision`, `Recall`, and `Accuracy` measurements.

Random Model

This model is used as a baseline model for comparison with the others.

When given an `event` (a tournament), it will randomly picky 3 participating teams.

The results are as follow:

Precision	Recall	Accuracy
0.1250	0.1533	0.1803

Volume Based Heuristic Model

This model was used to see how a simple heuristic approach would perform against a random model. This model aggregated the total volume sold of items relating to each team in a given tournament. It then picks the teams with the top 3 most volume of items sold. The items considered were any `Stickers` relating to the participating teams, even if they did not match the `Event`.

The results are as follow:

Precision	Recall	Accuracy
0.2125	0.2567	0.2043

Money Spent Based Heuristic Model

This was another simple heuristic model, this time calculating the average daily sales using $AVG(Price_i) * Volume_i$ where $i\in {Teams}$. At the same time, both this and the previous model give a good idea about the performance of self-constructed features.

The results are as follow:

Precision	Recall	Accuracy
0.2250	0.2600	0.2060

Volume Based Z-Score Random Forest

This model calculates each item's `z-score` for price per day, from 20 days before the event up until the end date of the event. It then groups these scores per team and calculates an average z-score per team. The underlying assumption is that when a team performs well, all `Stickers` relating to the team experience the same change in price.

The model is then used to predict the top 3 teams for each tournament, and became our best performing model.

The results are as follow:

Precision	Recall	Accuracy
0.7020	0.4650	0.8040

What Could Have Been Improved or Done Differently?

There are other features that we were unable to collect, mostly due to time contrainsts, that could have provided even stronger signal. Factors such as `Player/Team Highlights`, `Player Weapon Skins`, `Player Weapon Skins Sale Data`, `Bracket Seeds`, `Game Updates`, and other events. Some items have skewed sale data during tournament events due to these other factors. Having these could help distinguish between items who's sale data are impacted by the "hype" rather than by team performance.

The entire scope of the project could have been different. The Steam Market data is a rich dataset. We thought about switching scopes to predicting market trends based on tournament results, recommending weapon skins, or classifying profitable cases. This discussion happened too late into the project, and we were unable to proceed with any of these.

Overall, we are able to successfully classify the top teams in each tournament by using sale data. To some extent, this could be used in a live setting to predict the results of an ongoing tournament, as well as translating these results to changes in the Steam Market.