Amir Abaskanov, Christian Savastano, Edgar Campos, Harsh Johari, Jordan Tandjung

CS 370 Team Project Report

Streamlit Link: <https://valorantanalysis.streamlit.app/>

GitHub Link: https://github.com/amirabaskanov/ValorantPlayerTypeAnalysis

Introduction and Background:

The esports industry has been a massively booming industry. The recent Valorant Champions Tour (VCT) 2024 Masters: Madrid tournament reached a new viewership peak at 1.7 million peak viewers during the grand finals (Figure 1). Being part of the Sentinels esports team, we are looking to create the best roster to win the 2025 VCT season. Currently, there is no software to analyze Valorant players, so we are looking to create a way to assess player performance and establish the best new recruits for our team. A data-driven tool like this would be very valuable, as player contracts are each worth hundreds of thousands of dollars annually, potentially adding up to millions.

As shown in the Sentinels Balance Sheet for 2023 and 2022, the club is operating at about an 8-million-dollar loss each year, with operational costs totaling about 10 million annually (Figure 2). Because esports in general is still relatively new compared to other athletic disciplines, especially in the case of Valorant (since it was only released in 2020), many teams operate at a loss and are not very profitable to their owners. With the help of a new tool to improve roster making efficiency, the hope is that it will allow Sentinels to have a better chance at winning tournaments and championships, which will allow our team to increase in popularity. This in turn will help with attracting more sponsors, winning larger prize pools, attracting customers to our current sponsor companies, and promoting Valorant gaming to new audiences. All these expected effects will then also make it more worth it for investors and owners to stay with the team as it increases their trust and confidence in their investment.

Figure 1: VCT 2024: Masters Madrid viewership summary

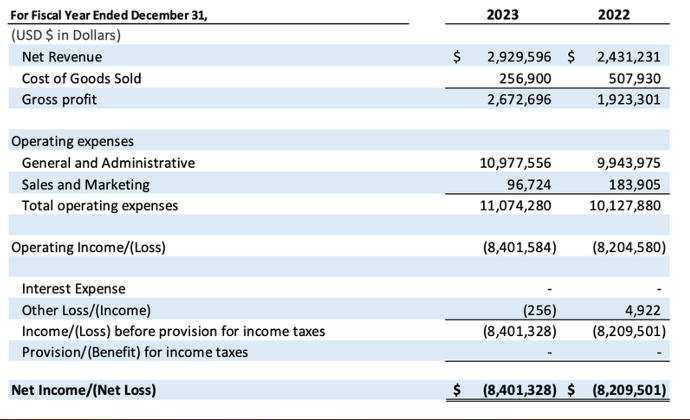


Figure 2: Sentinels Corporation Balance Sheet 2023 and 2022

Data and Methods:

Player and team data was collected through each game played during the 2024 VCT season and compiled on Kaggle. To best quantify the different playstyles and performance metrics of each player, we focused looked five key characteristics: Average assists per round, average combat score, clutch ratio, average number of first deaths per round, and headshot percentage.

Other statistics were explored, but generally made clustering techniques used weaker and unable to detect group differences easily. Statistics were chosen due to qualitative being a wholistic view of a player while being quantitatively distinct and having low correlations to each other. Other statistics considered included kill:death ratio, “kill, assist, trade, survive %”, average damage per round, kills per round, first kill per round, and rating. The different statistic combinations were reviewed manually using correlations and plotting players in 3d space with various statistic combinations. This process of reviewing variables showed the importance of picking representative variables and the care needed to avoid very similar variables: without this care, individuals within the dataset seem too similar to every other player for a model to create distinct clusters.

Player data was stored in a mix of ratios per game, strings, and numbers. Headshot percentage was stored as a string formatted “XX%”. This was converted to a number of .XX and treated as a ratio for the game. Ratios for the game included average assists per round, average combat score, average number of first deaths per round, and headshot percentage. These multiplied by the rounds in a game to ensure a high-round game properly had more influence on a player’s overall stats. Clutch ratio was stored as a string of “X/Y” where X is the successful clutches and Y is the potential clutch opportunities. The values were separated and added across all games without weighting based on rounds. Clutch ratio became an overall clutch success rate. If a player had no clutch opportunities, the success rate was marked as 0%. The variety of data formats presented a challenge, so we separated the data as needed to process each challenge separately, then combined the player data at the end. Player data was finally normalized using a 0-1 scale where the smallest value for a statistic is 0 and the largest is 1.

Once player data was normalized, a clustering algorithm was applied. While we assumed there would be 3-4 clusters, we did preliminary analysis using Agglomerate Clustering with various numbers of clusters to verify if our intuition was reasonable or if there was a better number of clusters than we thought. We found that using three clusters created the best distinct groups, with a silhouette score of .38, as shown in Figure 3.

After clustering, player data was combined with match data. This was necessary as player data contained a player’s statistics per round, but match data contained the outcome of the match. Using the data that overlapped between the datasets, unique MatchKeys were created and used to add a list of team players to each match. Matches were determined to be a win or loss by checking if the round win conditions summed to 13, the round victories needed to win a match. Players were converted to their cluster labels, and the player labels were counted to make an overall team composition. Once match win/loss and team composition were combined into one dataset, a statistical analysis was run on the relationship between winning and losing. We applied the clusters to team compositions, and we found that all cluster 0 players were part of Chinese teams and all Chinese teams were comprised of all five players being type 0 players, except for Edward Gaming which had four. This is likely due to differences in regional playstyle, a common occurrence in many multiplayer games due to separate servers between regions. This showed the importance of domain knowledge when analyzing data using machine learning.

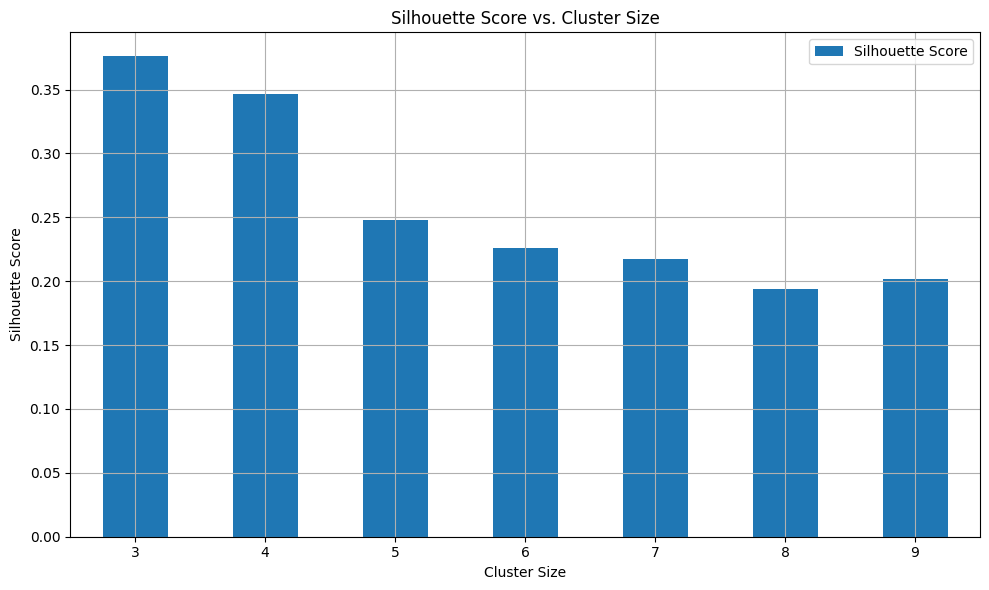


Figure 3: Bar graph showing agglomerate cluster size vs silhouette score

Results:

After creating the different clusters (Figure 4), we found that players in the VCT fall into three general categories: team players (type 0), defensive players (type 1), and aggressive players (type 2) based on domain knowledge and review of the statistics. There were 67 players that fell into the team players group, 52 in the aggressive group, and 141 in the defensive players group. When analyzing the optimal team composition, we found (4, 1, 0) had the highest win rate, but this represents only one team, Edward Gaming. Of common compositions, (0, 3, 2) had the highest win rate. Applying this to the business scenario: Tenz and Sacy are retiring from Sentinels this year and are both type 1. They should be replaced with a type 1 and 2 player to achieve the optimal combination of (0, 3, 2).

Improving this model comes in two major forms. The first is applying a time series consideration to see if teams improve when changing compositions or if the best combination changes depending on the other combinations present. Win rates changing when team composition changes could verify that there are generally better team compositions. Alternatively, if the best team composition depends on the surrounding team compositions, there may be a ‘rock/paper/scissors’ scenario where certain combinations are just better or worse against others.

The second is reviewing if players change based on the team they join, or if teams hire players who fit the playstyle. For example, the Chinese teams were generally comprised of all cluster 0 players, but many players were from other countries or even continents. This may indicate players adapt to the team they join, in which case another method of analyzing players may be more useful. It would also show that clustering may be a more useful tool for team analysis rather than player analysis.

Resources and Lessons:

We generally learn more about pandas’ features and data processing through this project. We also learned the importance of domain knowledge to analyzing a business problem. This highlights the importance of teamwork, as the strengths or knowledge of one member can cover the weaknesses of another. Using clustering also showed us the importance of picking relevant and distinct variables; without this, clustering algorithms will perform worse due to individuals appearing more similar.

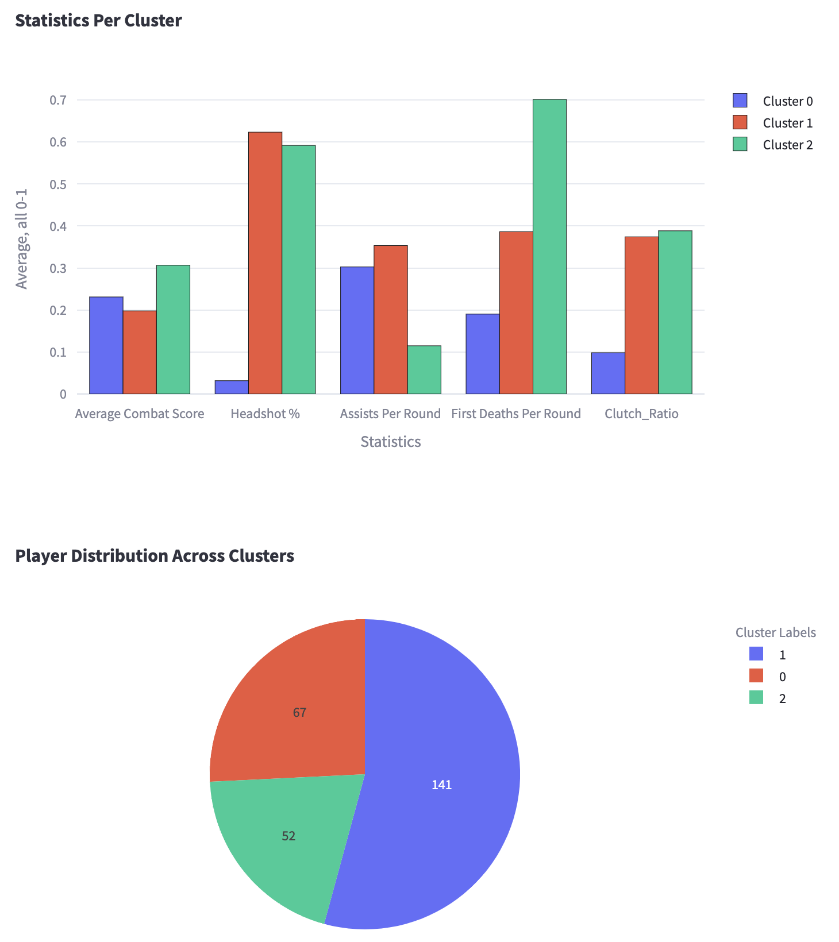


Figure 4: VCT Player normalized summary characteristics based on clusters

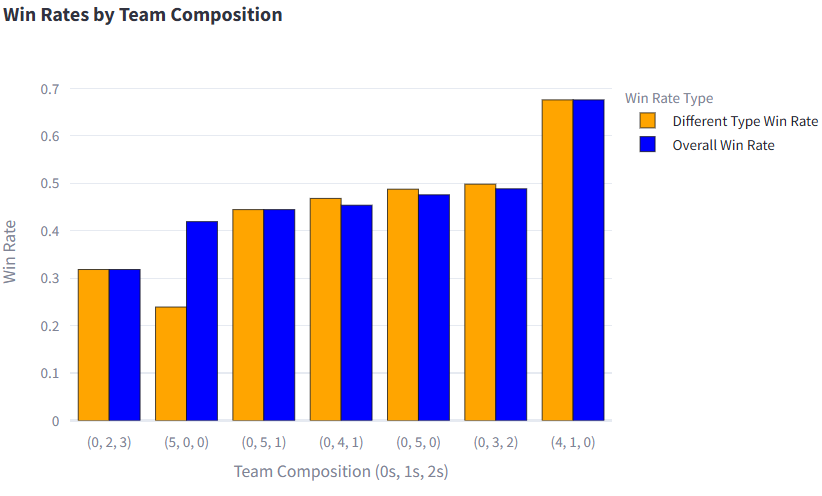


Figure 5: Bar chart showing win rates by team composition

Project Journey:

Meeting November 21: Deciding dataset to use. After reviewing our interests and the available datasets, we chose the 2021-2024 Valorant tournament dataset. We also decided to conduct an analysis on clustering players to identify the value of team composition.

Meeting December 2: Review of the dataset, our goals, and division of labor. Christian and Amir will primarily write code while Harsh, Jordan, and Edgar will primarily focus on making the presentation. All members will review the data to make assumptions about meaning.

Meeting December 3: Christian has written code for clustering. Christian quantitatively reviewed metrics for analysis. Edgar and Amir qualitatively reviewed metrics for analysis. Christian and Amir worked on various parts of result analysis. Edgar, Harsh, and Jordan created the presentation and made business application of current model.

Meeting December 4: All members met prior to class to review the presentation, practice, and adjust where needed.

The primary challenges involved choosing variables and interpreting results. Solving both challenges involved quantitative and qualitative reasoning. Picking variables meant choosing variables based on what generally defines different players, but we needed to reduce the variable selection by reviewing correlations and reduce the unnecessary variables. Without the former, the clusters would be coincidental rather than meaningful, but the latter was important for clusters to be generated. Interpreting results required reviewing the statistics for each player group and the game knowledge to understand how the results translated to real players. We also required game knowledge to understand why there was as clear regional difference between clusters.