Connor Broyles

Exploratory Data Analysis

Data Science Senior Capstone

2/2/2024

Cgoodman2@bellarmine.edu

Project Using Python, Tableau, and Microsoft Word

Introduction

Exploratory Data Analysis or also known as EDA is a crucial phase in the process of analyzing and understanding a dataset. It involves the initial examination of data to summarize its main characteristics, employing statistical graphics and other data visualization methods. The primary goal of this EDA is to gain insights into the underlying structure, patterns, and relationships within the Valorant dataset, which can help guide further analysis and hypothesis formulation. During this EDA, data analysts this project and analysis will explore various summary statistics, distribution plots, scatter plots, and other visualizations to identify patterns, outliers, and potential trends. This process is essential for making informed decisions about subsequent analyses, model building, and drawing meaningful conclusions from the dataset.

The dataset for this project is statistical data from the online video game named Valorant. Valorant is a first-person shooter online competitive game with a 5v5 style. There are 2 teams of 5 and they compete to be the first team to win by 13 rounds or “points”. Each team can kill the other and receive specific statistics in kills, deaths, first kills of the round, headshot percentage, etc. and these statistics are logged and reflected within the data set. The dataset can be found off of Kaggle and a direct link will be in resources at the end of the analysis.

I chose this dataset for the volume of statistics. The dataset consists of just under 250,000 instances with specific data on each teams kill, deaths, assists, maps played, characters chosen, rating, ct side rating, t side rating, average combat scores, and more. Some data like average combat score for example is a great way to see very specific ways of how people play. This will be great for developing models to get more accurate predictions and data. All of these statistics are stats used in game by each team and player. Even if someone does not understand the game specifically, they will be able to understand the statistics.

Data Set Description

Name	Data Type	Range of Values	NaN Percentage	Description
Match-datetime	Interval	2023 Match Date	0%	Day/Month/Year in which the match was held and data was taken from.
Patch	Nominal	6.00-6.06	0%	Game Patch – Patch refers to specific update that the game was played on. | I.E. patch 6.1 is the first update of Valorant of year 2023
Map	Nominal	Na	0%	Each game is played on a different environment in the game and we call those specific environments “Maps”.
Team1	Nominal	Na	0%	One team name of the match out of 2
Team2	Nominal	Na	0%	One team name of the match out of 2
Team1-score	Ratio	0-24	0%	Team 1 score report. The max number of rounds in Valorant are 13 unless they are tied which can go all the way up to 24. If they won the match, they will have a higher number then the other score team score.
Team2-score	Ratio	0-24	0%	Team 1 score report. The max number of rounds in Valorant are 13 unless they are tied which can go all the way up to 24. If they won the match, they will have a higher number then the other score team score.
Team1-win	Ratio	0-1	0%	Either a 0 for loss or 1 for win.
Team2-win	Ratio	0-1	0%	Either a 0 for loss or 1 for win.
Player-name	Nominal	Na	0%	In game username of the professional player that the stats were taken from.
Player-team	Nominal	Na	0%	In game team name of the professional team that the stats were taken from.

Agent	Nominal	Na	0%	Valorant consists of over 15 characters that have different abilities and things you can do in game. These characters are called agents and can be chosen by each individual player.
Rating	Ratio	0-3	18%	Specific in game rating score based off of play stats vs average characters in game.
Rating-t	Ratio	0-3	18.5%	Specific in game rating score based off of play stats vs average characters in game for “Terrorist Side”. Often called t-side but it simply means attacking side of the match. One team attacks and one team defends the defending side is nick named “Counter terrorist” or “Ct for short”.
Rating-ct	Ratio	0-3	18.5%	Specific in game rating score based off of play stats vs average characters in game for “Counter-Terrorist Side”. Often called ct-side but it simply means defending side of the match. One team attack and one team defend the defending side is nick named “Counter terrorist” or “Ct for short”.
Average combat score	Ratio	0-750	12%	Averaging statistic using stats like kills, headshot percent, hit count of bullets, and other stats all combined into one flat integer. This can range from 0 to almost 750 yet averages around 200-400 range. The higher the number the better overall the player did in the match. Very important stat!
Acs-t	Ratio	0-750	14%	Averaging statistic using stats like kills, headshot percent, hit count of bullets, and other stats all combined into one flat integer. This can range from 0 to almost 750 yet averages around 200-400 range. The higher the number the better overall the player did in the match. This is specifically for attacking side team.
Acs-ct	Ratio	0-750	47%	Averaging statistic using stats like kills, headshot percent, hit count of bullets, and other stats all combined into one flat integer. This can range from 0 to almost 750 yet averages around 200-400 range. The

				higher the number the better overall the player did in the match. This is specifically for defending side team.
Kills	Ratio	0-40	1%	Number of other players killed per match normally ranging from 0-20 but can reach up to almost 40.
Kills-t	Ratio	0-40	1%	Number of other players killed per match for attacking side normally ranging from 0-20 but can reach up to almost 40.
Kills-ct	Ratio	0-40	1%	Number of other players killed per match for defending side normally ranging from 0-20 but can reach up to almost 40.
Deaths	Ratio	0-25	12	Number of times the player died per match. A player can only die once per round and normally there's on average 15-20 rounds per game.
d-t	Ratio	0-25	13%	Number of times the player died per match for the attacking side. A player can only die once per round and normally there's on average 15-20 rounds per game.
d-ct	Ratio	0-25	13%	Number of times the player died per match for defending side. A player can only die once per round and normally there's on average 15-20 rounds per game.
Assists	Ratio	0-50	13%	Assists are basically if someone on the enemy team dies and you helped in some way, maybe flashbangs, healing allies, and other helpful ways to help your team are counted by this stats.
a-t	Ratio	0-50	13%	Assists for attacking side. Assists are basically if someone on the enemy team dies and you helped in some way, maybe flashbangs, healing allies, and other helpful ways to help your team are counted by this stats.
a-ct	Ratio	0-50	13%	Assists for defending side. Assists are basically if someone on the enemy team dies and you helped in some way, maybe flashbangs,

				healing allies, and other helpful ways to help your team are counted by this stat.
Total kills minus deaths	Ratio	-25-25	13%	Total amount of kills the players obtained over the match minus their total number of deaths. This can result in a negative number if they received more deaths than kills in a match.
Tkmd-t	Ratio	-25-25	23%	Total amount of kills the players obtained over the match minus their total number of deaths for attacking side. This can result in a negative number if they received more deaths than kills in a match.
Tkmd-ct	Ratio	-25-25	21%	Total amount of kills the players obtained over the match minus their total number of deaths for defending side. This can result in a negative number if they received more deaths than kills in a match.
Kills assists survive trade %	Ratio	0-1	14%	A 1 number percentage statistic including kills, assists, and the trade percentage. Trade percentage is if they killed another player before they died. This is important because if everyone kills a player before they die per round, they win the round due to number advantage.
Kast-t	Ratio	0-1	14%	A 1 number percentage statistic including kills, assists, and the trade percentage for the attacking side. Trade percentage is if they killed another player before they died. This is important because if everyone kills a player before they die per round, they win the round due to number advantage.
Kast-ct	Ratio	0-1	70%	A 1 number percentage statistic including kills, assists, and the trade percentage for defending side. Trade percentage is if they killed another player before they died. This is important because if everyone kills a player before they die per round, they win the round due to number advantage.

Average damage per round	Ratio	0-800	11%	A player has 100-150 health points in game. Getting shot reduces these hit points and if their heath reduced to 0 the player dies. This stat shows how much damage the player did per round. Normally ranging from 0-200 but can get up to almost 800.
Adr-t	Ratio	0-500	14%	Damage dealt to players. This stat is only for attacking side.
Adr-ct	Ratio	0-500	65%	Damage dealt to players. This stat is only for defending side.
Headshot percentage	Ratio	0-1	13%	A player can shoot another player in the legs, chest, or head region of their player models. This percentage is only for if the player gets the kill if that player got the kill by shooting another player in the head.
Hs-t	Ratio	0-1	71%	Headshot percentage for attacking side.
Hs-ct	Ratio	0-1	69%	Headshot percentage for defending side.
First kill	Ratio	0-15	13%	First kills or “First Bloods” are important because it provides a number advantage to the specific team. First bloods are just the very first kill of the round and the teams that get higher of them have a higher likelihood of winning the round.
Fk-t	Ratio	0-15	13%	First blood for attacking side. First kills or “First Bloods” are important because it provides a number advantage to the specific team. First bloods are just the very first kill of the round and the teams that get higher of them have a higher likelihood of winning the round.
Fk-ct	Ratio	0-15	13%	First blood for defending side. First kills or “First Bloods” are important because it provides a number advantage to the specific team. First bloods are just the very first kill of the round and the teams that get higher of them have a higher likelihood of winning the round.

First death	Ratio	0-10	13%	First death is simply the first death a team receives. These are important because if a team has a high amount of first deaths, they are less likely to win the round i.e. the game.
Fd-t	Ratio	0-10	13%	First deaths for attacking side. First death is simply the first death a team receives. These are important because if a team has a high amount of first deaths, they are less likely to win the round i.e. the game.
Fd-ct	Ratio	0-10	13%	First death for defending side. First death is simply the first death a team receives. These are important because if a team has a high amount of first deaths, they are less likely to win the round i.e. the game.
First kills minus first deaths	Ratio	-15-15	13%	Total amount of first bloods minus first deaths. Can be negative if the amount of first deaths received are higher than amount of first kills received.
Fkmd-t	Ratio	-15-15	18%	Total amount of first bloods minus first deaths. Can be negative if the amount of first deaths received are higher than amount of first kills received. (Stat only for attacking side).
Fkmd-ct	Ratio	-15-15	24%	Total amount of first bloods minus first deaths. Can be negative if the amount of first deaths received are higher than amount of first kills received. (Stat only for defending side).

Data Set Summary Statistics

For the dataset summary statistics in this project, I will be using Python in Visual Studio code. The file can be accessed via the same repository on GitHub and be an "ipyn" file. The use of Python and its libraries as well as GitHub Co-Pilot was used in aid of the data analysis.

Some example photos to gauge what you might expect from the file include but are not limited to:

```
df = pd.read_csv("valorant_data.csv")
# Link to valorant dataset
# https://www.kaggle.com/datasets/qualidea1217/valorant-pro-matches-since-april-2021
```

```
num_instances = df.shape[0]
print("Number of instances:", num_instances)
```

✓ 0.0s

```
Number of instances: 249710
```

```
df.head()
```

✓ 0.0s

	Unnamed: 0	match-datetime	patch	map	team1	team2	team1-score	team2-score	Team 1 Win	Team 2 Win	...	hs-ct	first kill	fk-t	fk-ct	first death	fd-t	fd-ct	fi
0	0	4/16/2023 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	1	0	...	0.26	1.0	0.0	1.0	1.0	0.0	1.0	
1	1	4/16/2023 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	1	0	...	0.38	0.0	0.0	0.0	1.0	1.0	0.0	
2	2	4/16/2023 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	1	0	...	0.27	2.0	2.0	0.0	1.0	1.0	0.0	
3	3	4/16/2023 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	1	0	...	0.24	7.0	3.0	4.0	5.0	5.0	0.0	
4	4	4/16/2023 10:00	6.06	Haven	Impulse GW	EGN Esports	13	6	1	0	...	0.80	0.0	0.0	0.0	1.0	0.0	1.0	

5 rows × 49 columns

```
df.tail()
```

✓ 0.0s

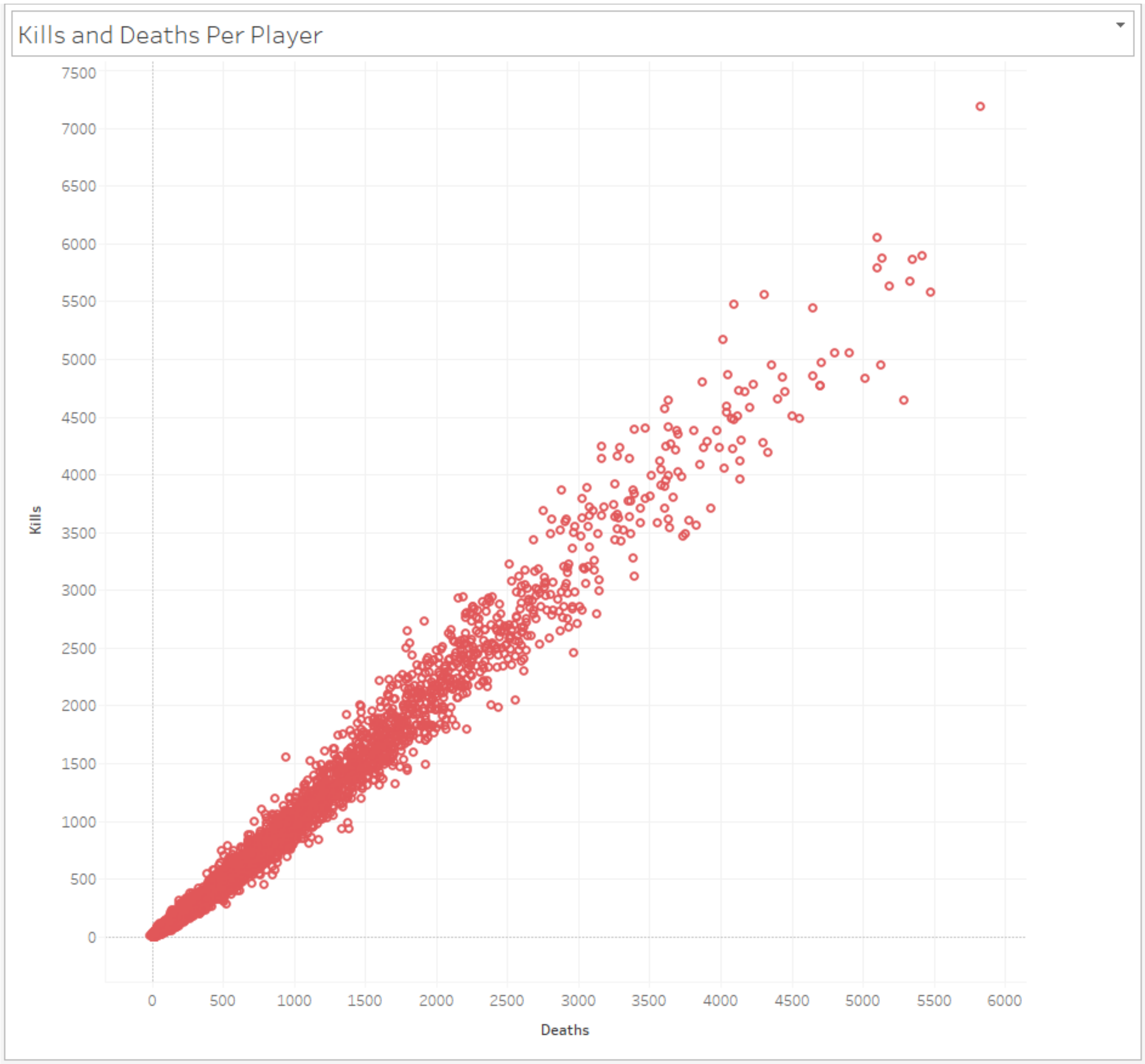
	Unnamed: 0	match-datetime	patch	map	team1	team2	team1-score	team2-score	Team 1 Win	Team 2 Win	...	hs-ct	first kill	fk-t	fk-ct	first death	fd-t	fd-ct
249705	249705	4/14/2021 5:30	2.07	Bind	Mindfreak	Team Bliss	14	12	1	0	...	0.38	2.0	2.0	0.0	2.0	2.0	0.0
249706	249706	4/14/2021 5:30	2.07	Bind	Mindfreak	Team Bliss	14	12	1	0	...	0.21	4.0	3.0	1.0	5.0	1.0	4.0
249707	249707	4/14/2021 5:30	2.07	Bind	Mindfreak	Team Bliss	14	12	1	0	...	0.13	4.0	1.0	3.0	3.0	1.0	2.0
249708	249708	4/14/2021 5:30	2.07	Bind	Mindfreak	Team Bliss	14	12	1	0	...	0.44	2.0	0.0	2.0	3.0	2.0	1.0
249709	249709	4/14/2021 5:30	2.07	Bind	Mindfreak	Team Bliss	14	12	1	0	...	0.18	0.0	0.0	0.0	1.0	1.0	0.0

5 rows × 49 columns

(And much more)

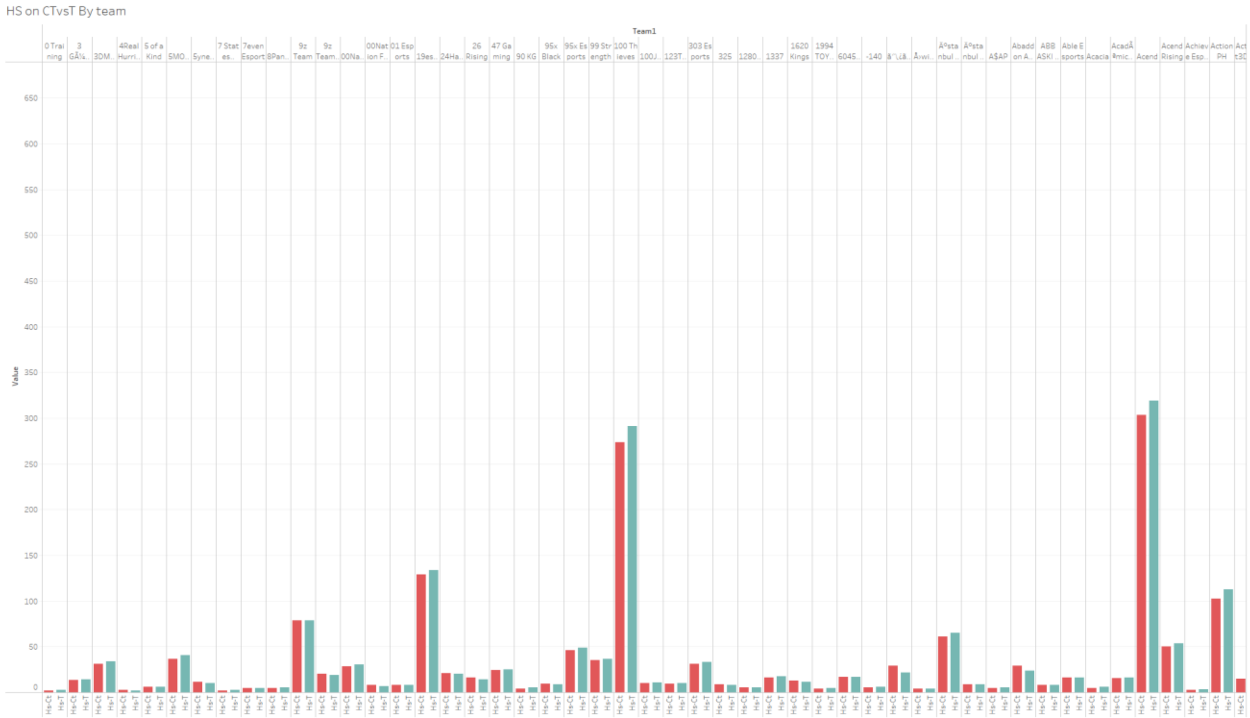
Data Set Graphical Exploration

All visualizations are done via the Tableau software. The file for visualizations for the EDA can be accessed from the GitHub repository as well as all the charts and data visualizations. When looking at this data there are key statistics that stand out and Tableau can help visualize those. For example:

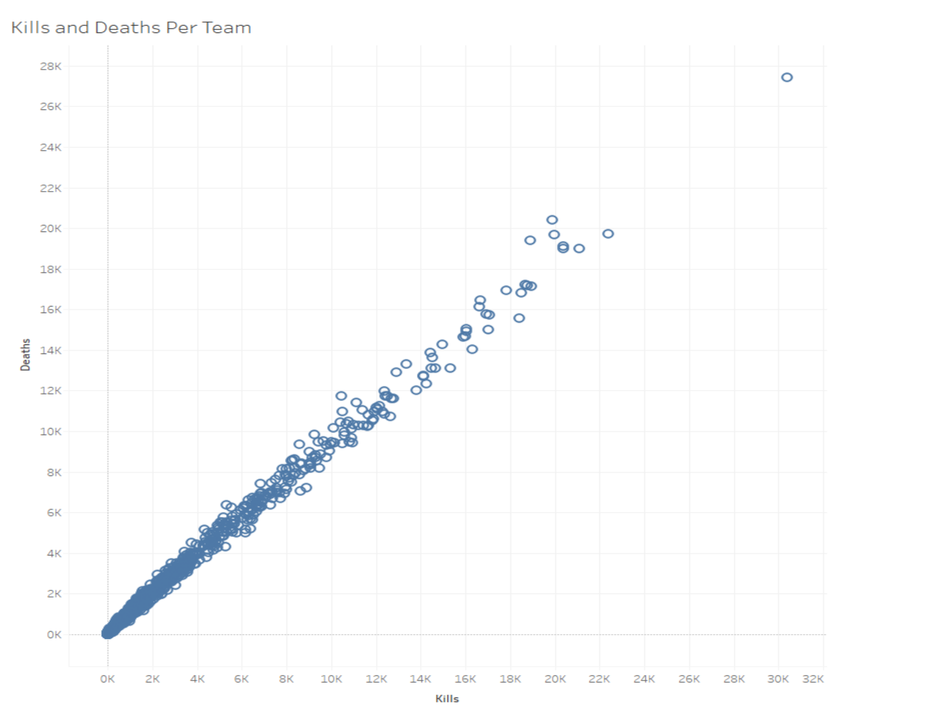


Above is the kills and deaths each player in professional play has. For example the dot all the way at the top shows us that "Reduxx" has the highest amount of deaths coming in at 5,825 while also having the highest amount of kills at 7,184. These statistics are important for how the game is actually played but also for teams stats as well to know who has a higher chance of getting a kill in game.

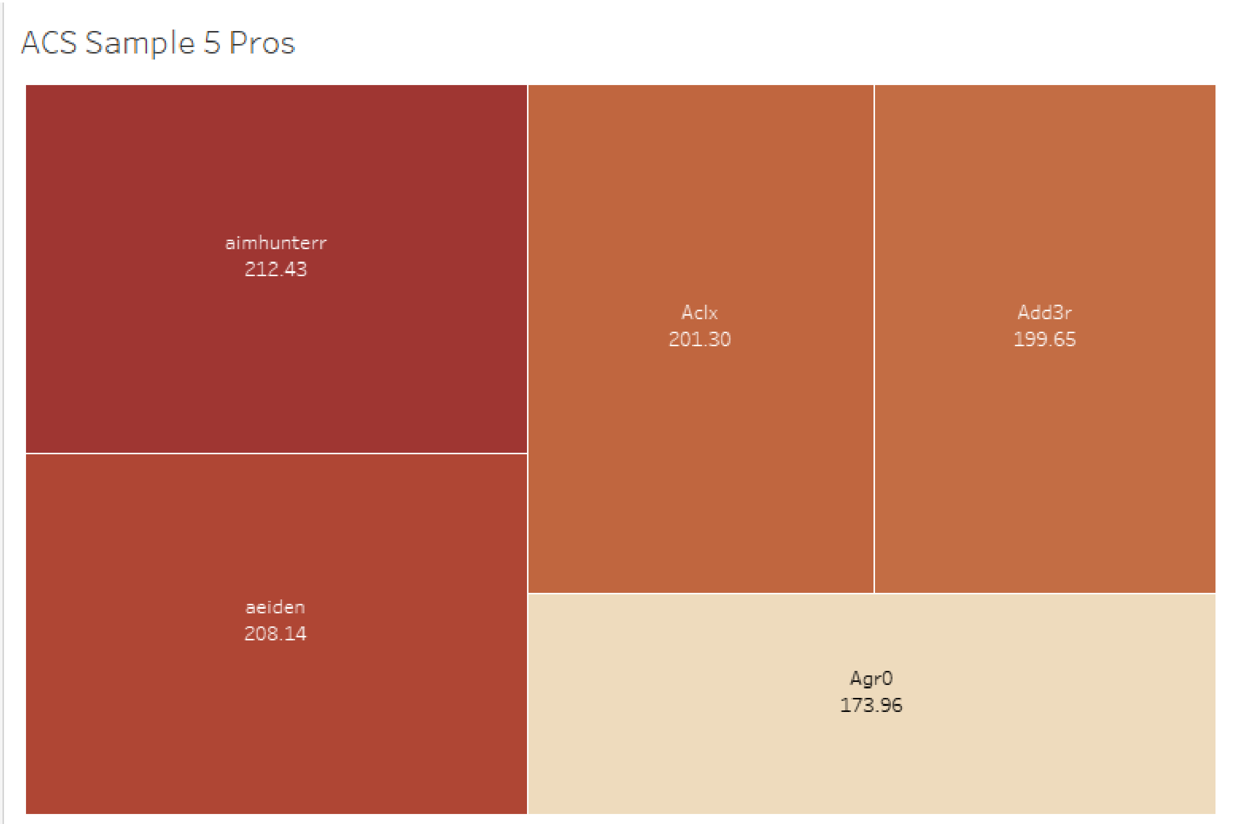
Another example of a few of the visualizations are of team Kills and death and headshot numbers by team:



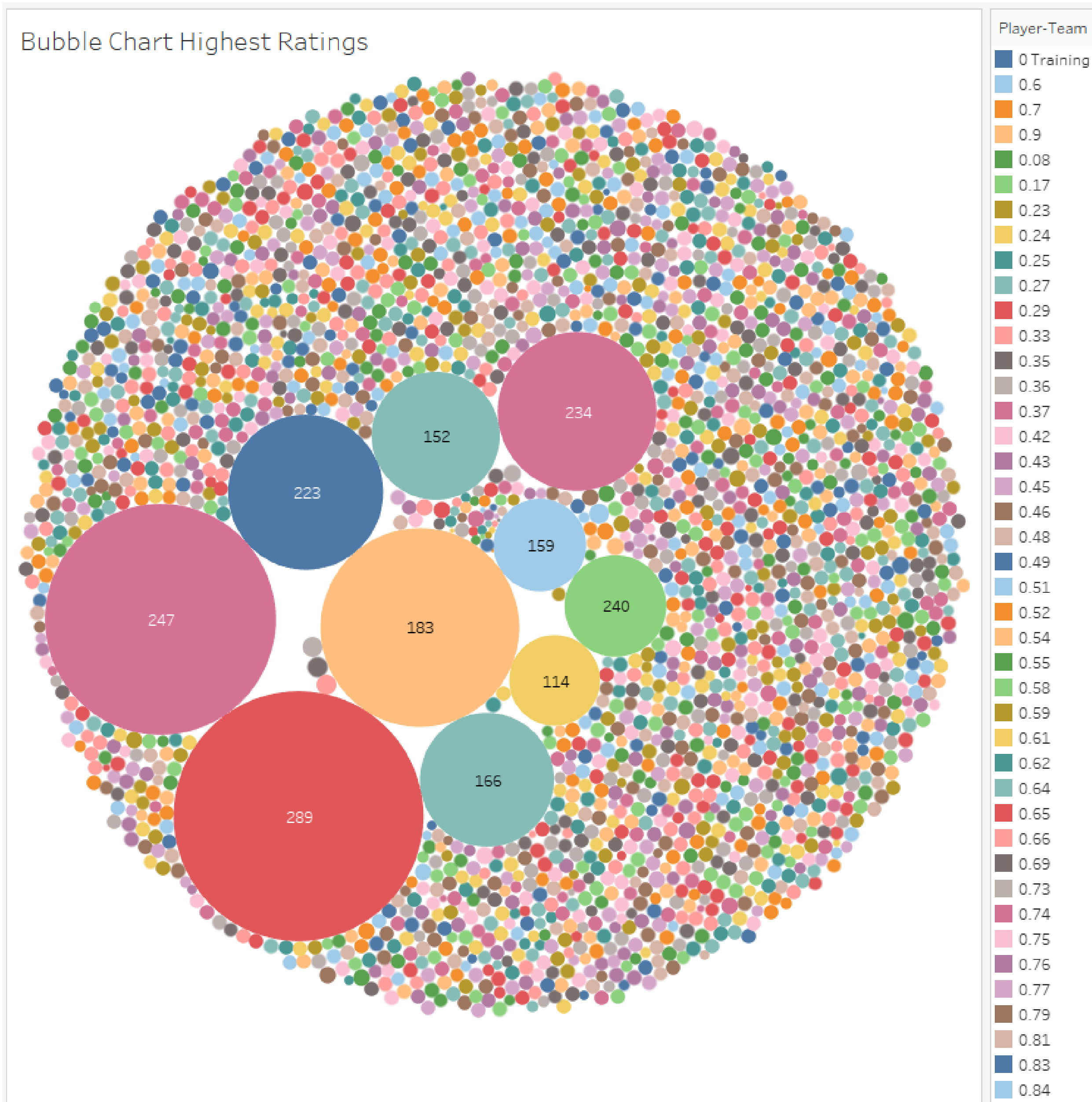
Above you can see how each team is categorized and then they have a number for their CT side and T side with regards to headshots.



The blue scatter plot is the summation of kills and deaths sorted by teams. For example, Cloud9 professional team might have a higher plot then lets say Team liquid for last year.



Above you can see 5 well known pros and their combat scores. We notice that Agr0 has a much lower average combat score then aimhunterr. I can sort their combat scores and which players I want in any way we choose via Tableau. Other ways of looking at combat scores to sample our data can be in the form of a bubble chart seeing what the largest average combat scores look like.



There are more data visualizations to be found from sampling the data in the Tableau file uploaded to the GitHub repository.

Summary of Findings

Throughout the exploratory data analysis there are key elements of the dataset that stood out. For example, in game statistics that matter more than others might include: kills, deaths, average combat scores, and first kills. Taking note of these variables is exceptionally helpful due to the fact that I can monitor them and see if there is enough data included which lucky there is, and take note for them to be used in the future. On the other hand, throughout the python data analysis variables that stood out to me where; acs-ct, and a few other ct or t variables that had over 45% missing values in their columns and some even reached up to 90%. These variables I will have to do more analysis on but I may have to drop some.

Other variables are incredibly helpful when making specific visualizations like map choice and agent choice but they probably will not be very helpful at all in conducting and building my model so I may have to drop those along with the 90% missing value variables as well. This will have to be explored with more detail later on when building the model but those are the initial assumptions. However, with the large sum of data I have hopeful and optimistic that for my project I will have ample data when it comes to building models.

For the other variables most have almost all values within the columns but some variables did have 5-15% missing values. At the moment I mostly plan to replace the missing values with a mean in python. A simple function can add this in and replace the missing values in no time. Although for some variables it is possible that I might have to replace them with 1s or 0s if the variables are more binary focused but it does not appear that will be the case for most of the variables. Other issues that might arise is some multicollinearity within the variables and I will have to figure out which variables make the most difference. At the moment I plan to simply make a couple different models and test their significance and see which are the most useful within those respected models.

Overall, I am very happy with the data and the small issues that I do have I am planning out the fixes already and can proceed with the project. There will be ample data for my models and will allow for a great exploratory project.

Resources:

Link to dataset off of Kaggle: https://www.kaggle.com/datasets/qualidea1217/valorant-pro-matches-since-april-2021

GitHub

GitHub Co-Pilot used for aid in coding syntax with Python and Jupytr Notebook

Tableau

Microsoft Excel

Python and Python Libraries

ChatGPT used in making outline for this EDA