**General Notes:**

1. GroupID and MatchID seem useless for our prediction model.
2. Might be worthwhile to split the data into squads, duos, and solos. We could create and train a model for each category, this may help with issues like a player being carried.
3. It may also be worthwhile to scrub outliers as there are bound to be some players who are being carried by friends. These players will likely have low kills and even in some cases die early but still win due to their friends winning. We could reduce this issue by removing outliers and splitting the groups up into different models. It’s impossible to be carried in solos and therefore we don’t need to worry as much about this issue with that model.
4. This may not be an issue as it seems GroupID being a thing seems to suggest the kills would be the total number of kills made by the group if that’s the case disregard 3.
5. I still think splitting the data by match type could be helpful as we could specialize each model and get better prediction results. (ie teamkills is a totally useless feature in a solo match.)
6. We need to decide what makes a group or player seem suspicious on a botting or cheating front; this may cause outliers; some bots are programmed to hide in bushes until they are the last of two groups and then score headshots until they win using aimbot. This could cause a player who has done nothing but gotten a single headshot winning which would affect our model.
7. Let’s make sure to keep our code well documented with Markdown cells if we’re using Jupyter. I think that will help a lot with understanding as the data is passed from one person to another for different tasks.

**Data Preparation Notes:**

1. Need to check business moment decisions to check for need for standardization or normalization. Gut feeling standardization is needed at the very least.
2. We need to check for linearity between x values and y and between x values. (Once we have these numbers we can discuss how we want to handle collinearity.)
3. We want to check for outliers in this stage and decide how to handle them. (scatterplot, boxplot, histogram)
4. We need to check for missing data and decide how to handle any columns with lots of missing data or rows.

**Modeling Notes:**

1. As said before I think we should make three models for winning predictions, one for each match type.
2. Model will need to be a regression model since it has a continuous and not discrete.
3. May be a worthwhile addition to make a model to catch bots. This isn’t necessarily in the project requirements, but could look good on our resumes since this is a real thing a company would need to keep their game free of cheating.
4. I can’t remember if this was something I learned in the past or something we’ve already covered together in class but anomaly detection could be used to catch outliers and/or cheating in our dataset.
5. To whoever handles model building it may be worthwhile to look in Cross Validation so the model accuracy is as optimal as possible.