Capstone Exploratory Data Analysis

Objective

To utilize exploratory data analysis and inferential statistics to gain insights on my capstone dataset.

Analysis

1. Are there variables that are particularly significant in terms of explaining the answer to your project question?

The underlying questions for my project are what variables positively correlate with winning a match Players Unknown Battleground. Winning in this game means being the last player / team standing. To increase threshold of victory we can consider it to be within the top 5 team placement.

In my previous visualizations I've found that a player doing more damage and getting more kills is the strongest variable correlating with winning a match

2. Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

My mentor has told me that Pearson Correlation values above .4 or below -.4 are to be considered significant, anything in between those values are too close to 0, are thuoght to be weak correlations.

When examining the correlations below, its important to remember when looking at the team placement variable that a negative correlation is a good thing, since the lower the team placement number translates into closer to victory (first place).

Some noteworthy strong correlations are the following (rough estimates based off of all three samples):

1. Damage & Distance walked: .4   (independent)
2. Distance walked & team placement: -.8   (dependent)
3. Distance driven & team placement: -.5   (dependent)
4. Distance driven & survival time: .6   (dependent)
5. Player Kills & team placement : -.4   (independent)

I say player kills and team placement are independent because at the end of the day to win a match you don't have to get any kills or you may likely need only one kill. I don't believe these variables are strongly dependent on each other.

3. <u>What are the most appropriate tests to use to analyse these relationships</u>

I conducted a couple t tests using the scipy module, ttest_ind, giving me the test statistic and p value for two variables in the dataframe. I tried the test on two different variables, both times with extremely high test statistics and extremely low p values.