Milestone Report Capstone 1 PUBG

1. Problem statement: Why it's a useful question to answer and for whom (get this from your proposal)
2. Description of the dataset, how you obtained, cleaned, and wrangled it (get this from your data wrangling report)
3. Initial findings from exploratory analysis (get this from your data story and inferential statistics reports)
    1. Summary of findings
    2. Visuals and statistics to support findings

Problem Statement

Players Unknown Battleground became the top game of 2017 worldwide. The battle royale genre of video games can be very competitive, and only one player / team can be the winner. There are many different approaches to trying to win a match, the bottom line comes to caution vs aggressiveness. So what correlations from other variables can we find, that first place consistently has. These answers can help create strategy guides for PUBG, and potentially other battle royale games, which have dominated the gaming industry in the past two years.

Dataset

I obtained the dataset from Kaggle. The dataset didn't have any missing entries. Fortunately due to the nature of getting the data from a video game, having a missing value is not likely to ever happen. The dataset came in 5 csv files each around 2 gb. The columns included player distance walked, player damage, team placement, distance driven, survival time, and party size.

To clean the data I narrowed it down to the useful columns. I removed undesirable rows such as players that did not survive for a minute, or didn't travel a distance of more than a few meters. I

divided the player survival time column by 60 to convert it into minutes, making any graphs I conduct into a more discrete integer.

Most importantly I divided the data by game modes. The game can be played in solo's, duo's, or in a squad (up to 4 players on a team). Not mixing this data is essential. I received samples of 1000 for each game type to reduce the file into one small enough to do computations easily.

Inferential Statistics

I utilized seaborn joint plots to observe pearson correlations along with p values of multiple pairs of variables. I picked which variables to compare by first using the corr() method on each file, and observing which pairs of variables contained correlations greater than .4 or below -.4. From this I could clearly see that that first place, along with the top 5 players produced more damage and kills than those placed lower. Killing and doing damage is not necessary to win a match, but the pearson correlations and low p values showed that they consistently were high for first place.

To further demonstrate this, I took from the complete dataset for solo matches all the players that won first place, took bootstrap samples 1000 times, and got the confidence intervals for the means. I did the same for second place. First place consistently has a mean kill of about 7, with a confidence interval of [6.94742282, 6.99468351], while second place had a consistent mean of about 3.6, and a confidence interval of [3.58724136, 3.61285905].