Title: PlayersUnknown's Battleground's Data Royale

Team members: Thomas Alder, Federico Aragon, Matthew Davenport, Hao Dong

Description: 2-3 sentence paragraph project description - what interesting questions do you intend to answer?

In this project we are looking to analyze data from the results of a player's matches they played in a Battle Royale video game named PlayerUnknown's Battlegrounds. Using this dataset, we hope to answer questions regarding the game's mechanisms through an analysis of of player damage and deaths and how this statistic could be improved. We would also like to see what the likelihood of a player being a hacker is through understanding the inflicted damage they cause and whether or not there are significant outliers which may suggest an unfair advantage. Lastly we would like to find the best strategy to use when playing the game in terms of determining the most dangerous places to avoid, places which may be beneficial, and the general strategy of players that tend to win more.

Prior Work: What prior work has been done on your idea

Currently on Kaggle there is a competition which uses historical data from the game to predict player's finishing places in the match. Work has also been done with data visualization such as heat maps representing the deaths throughout the games and the heavily visited areas on map. In addition, work has been done analyzing the distance of kills, analysis on survival times, and further predictions on the locations of circles and player drops.

Datasets:
- List of datasets to use
  - https://www.kaggle.com/skihikingkevin/pubg-match-deaths/home
  - https://www.kaggle.com/c/pubg-finish-placement-prediction/data
- Where found (URL and who is supplying the data, e.g., NASA)
  - Found via Kaggle, supplied by users skihikingkevin and Kaggle.
- Whether you have it downloaded (on who's machine)
  - The dataset is downloaded on Thomas's and Matthew's machine.

Proposed work: what do you need to do?
- Data cleaning:
  - Shorter games in the dataset which have missing data must be removed.
  - Longer games that have few missing values will have those values filled with placement values.
  - If the number of players or teams in a game is too small then remove this data as well as it might skew the information we want to gain.
- Data preprocessing:
  - Reducing the number of attributes through correlation analysis.

- ■ Removing columns which have no association with match statistics like player_name and team_id.
- ■ Transforming data when necessary, for example if percentages are stored as integers or floats greater than 1, then convert them to their corresponding decimal values.
- ■ Normalization of attributes.
    - ○ Data integration:
        - ■ Merging the two separate datasets provided in the PUBG Match Deaths and Statistics Kaggle page.
    - ○ Etc.
        - ■ Data mining - analyze the data to gain useful knowledge from it, and use this knowledge to answer our initial questions.
        - ■ Data visualization - presenting the information found in a manner that is understandable by players of the game and non-players alike.

List of tool(s) you intend to use
- ○ Python - clean and integrate data, data visualizations, etc.
- ○ SQL (MySQL or Postgres + DBMS) - only if necessary for Data Warehousing
- ○ Google Docs - project reports
- ○ GitHub - project repo, and milestones

Evaluation: How you can evaluate your results?

Ideally with the correct results we would be able to create prediction scores that will be used to predict winners for a given match and players statistics. We would also like to predict match results given certain scenarios and environmental factors within a game such landing locations, total number of teams, equipment picked up, etc. These factors can be displayed on heat map data visualizations which portray "hot spots" around the two different maps of the game, scores can be given to different areas regarding their "dangerous" levels.