**DSCI-549: Introduction to**

**USC** Viterbi School of Engineering

**Computational Thinking and**

**Data Science**

**Term: Fall 2021**

**Project – Part 1**

**Date due: FRIDAY, November 5th 2021 - 5pm PT**

**Goals**

The purpose of the project is to familiarize yourself with the steps involved in most data science projects, at least in the conceptualization phase.

**Overview**

Choose a data science project of interest and answer the questions in the detailed assignment section below. Answer each question individually. **DO NOT** submit an essay or report-like assignment (you will be marked down).

Remember that **ALL ANSWERS MUST BE JUSTIFIED and MUST APPLY TO YOUR DATASET.** e.g., when describing the algorithm, ensure that you explain which columns/image… in the dataset the analysis will be applied to. Even though you are not writing code, you are expected to provide a close recipe.

You are allowed to work in groups. **HOWEVER,** this is an individual assignment, and your data science question/dataset/analysis must be unique. See the Course Overview slides for examples.

This is the first part of the assignment. You are allowed to change you question/dataset for the second part; in which case you will have to redo this assignment for context. Note that it will not be regraded.

**Assignment**

Part 1: Data (50 points)

1. Define an interesting problem that could be answered through data analysis. Describe the problem to a non-informed audience (10 points)

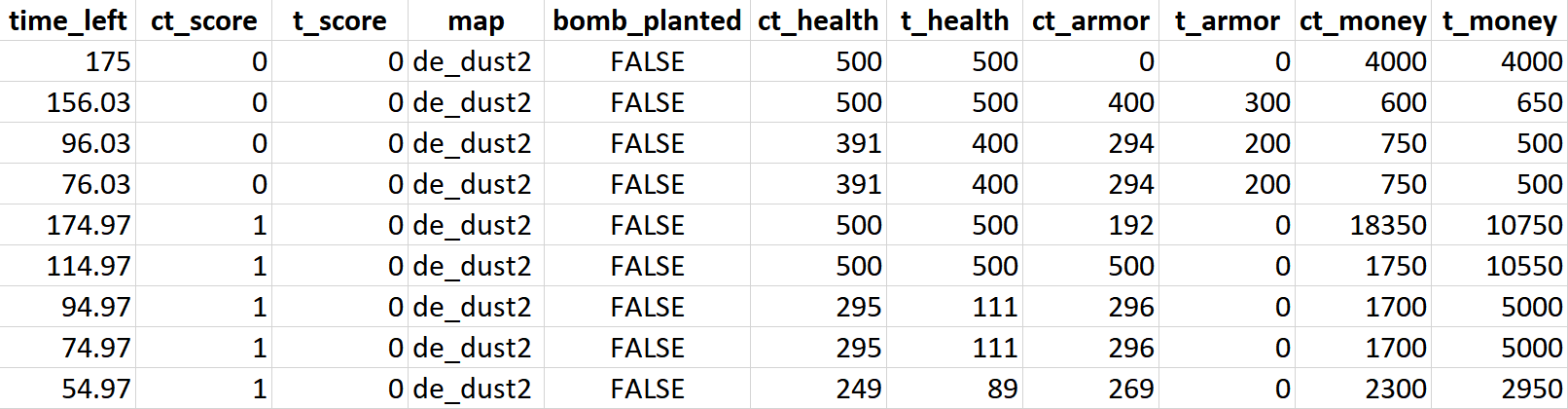
Counter-Strike: Global Offensive (CS:GO) is an multiplayer online shooting game developed by Valve. There are multiple game modes in CS:GO, however, the most prominent and widely played mode is called Competitive, in which two teams, also known as Terrorists (T) and Counter-Terrorists (CT), of 5 members will play against each other for a maximum of 30 rounds. The Terrorists’ objective each round is to plant a bomb or eliminate all the Counter-Terrorists, while the Counter-Terrorists’ goal is to defuse the planted bomb or eliminate all the Terrorists.

During gameplay, there are various factors that will help each team gain their advantages towards victory. For instance, the amount of armor, the amount of firepower (pistols, riffles, grenades), the number of players alive, or the number of defuse kits. The goal of this project is to take a look into those factors and determine which ones have the most influence of the outcome of the round, or even more specific, if given all the variables at a moment in the round, can we predict which team will come out on top?

1. Find one or more datasets on the Web that are relevant to these questions, and that are: 1) accessible, 2) released with an open license, and 3) machine processable. For one of the datasets, specify:
   1. A brief description of what the data represents, with a small excerpt of the data as an example (10 points)

The dataset used in this project is “CS:GO Round Winner Classification” by Christian Lillelund. Below is the link to the dataset on Kaggle: <https://www.kaggle.com/christianlillelund/csgo-round-winner-classification>

a. The dataset contains snapshots of around 700 Competitive CS:GO games in 2019 and 2020. The snapshots are recorded 20 second between each other until the round ends. There are a total of 122411 snapshots, corresponding to 122411 rows and 97 features (The weapons that the Terrorists have, the amount of armor the Counter-Terrorists have, etc.), corressponding to 97 columns. Below is an example of the first 11 features of some snapshots in the dataset:



Take row 1 for example. Firstly, there are 175 seconds left in the round and the game has just begun since CT score and T score are both 0. The teams are playing in a map named de\_dust2, and looking at the other variables, it looks like that the round has just begun because all the values of the features are default values.

* 1. A discussion on how the dataset satisfies the three criteria above (20 points)

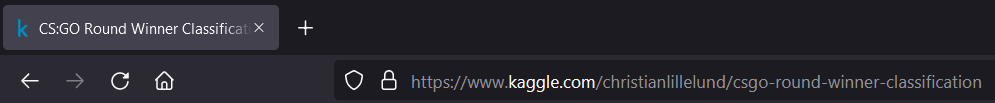
The dataset satisfies the requirements of the question because:

1. It is accessible and downloadable from Kaggle
2. Its license is CC0: Public Domain
3. It can be processed by machine because it is a Comma Separated Value (csv) file, is tabular data that can be read by many softwares
   1. The **details** for how you can access the data (10 points)

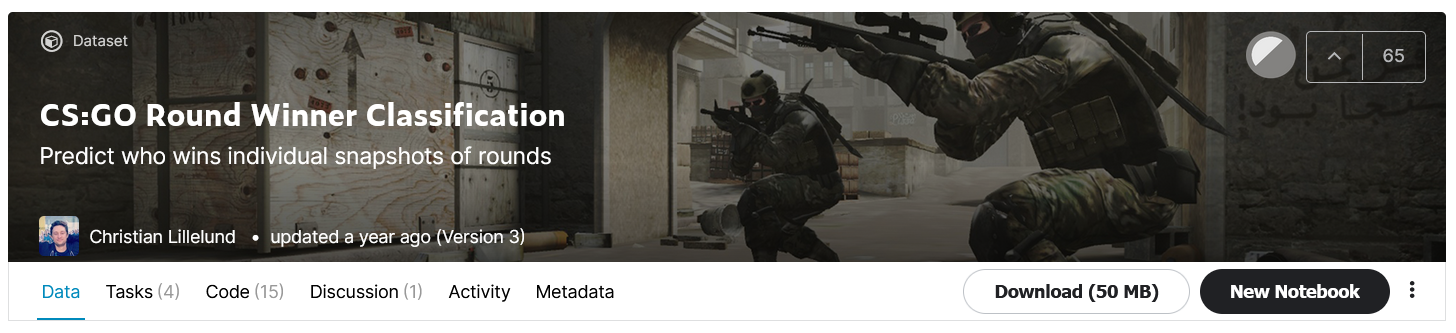
**Use** the technical terms that you have learned in class to describe data.

To access the data, follow the following steps:

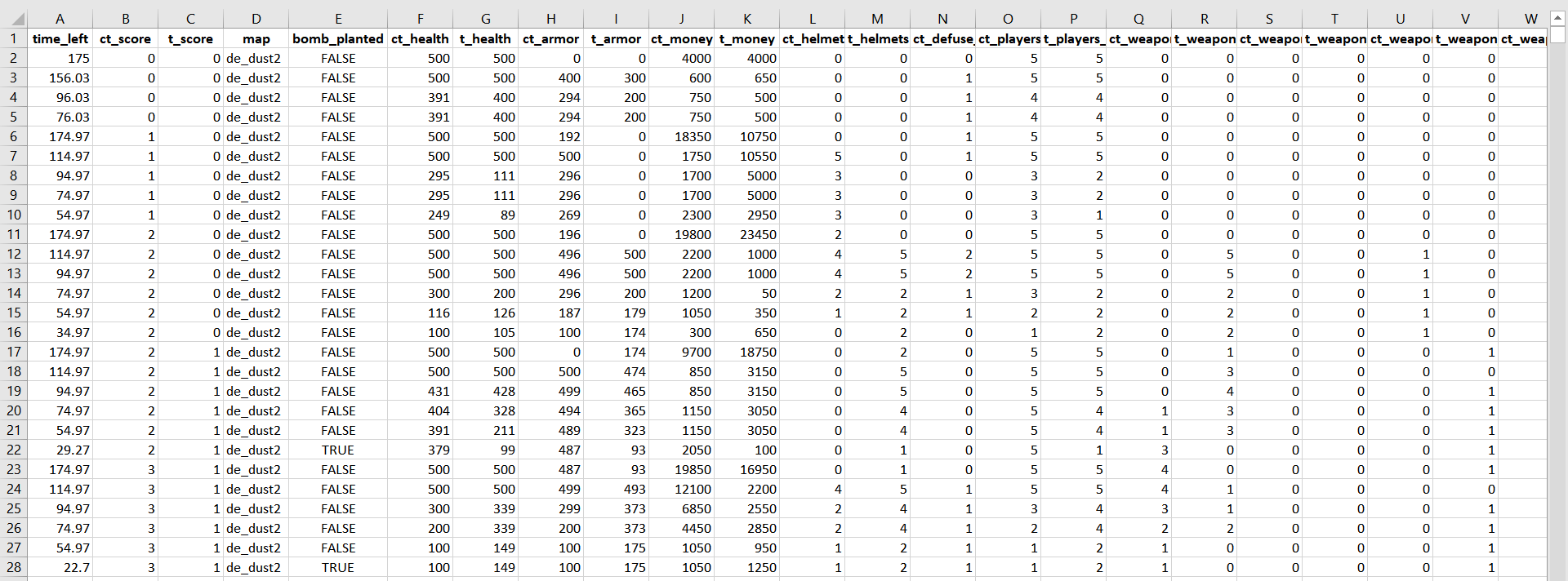
* Open browser and go to : <https://www.kaggle.com/christianlillelund/csgo-round-winner-classification>



* Click the Download (50 MB) button to download the dataset to your computer



* Open the file with Excel or other softwares



Part 2: Hypothesis (30 points)

1. Is the study you want to conduct experimental or observational? Why? How does that affect your conclusions? (10 points)

The study is observational because we simply measure or survey the data without affecting the data itself. In other words, the study does not intefere with the data and only observe what is happening with it. The fact that this is an observational study makes it impossible to establish causual connections and only allows us to associate variables.

1. State the null and alternate hypothesis (10 points)

H0 (Null hypothesis): There is no statistically significant evidence between the different factors of a CS:GO match and its winner.

H1 (Alternate hypothesis): There is statistically significant evidence between the different factors of a CS:GO match and its winner.

1. What are the dependent and independent variables? (10 points)

The independent variables are variables that we believe affect other variables and are manipulated in a study. Therefore the independent variables in this study are the 96 features (columns) of the data:

['time\_left', 'ct\_score', 't\_score', 'map', 'bomb\_planted', 'ct\_health', 't\_health', 'ct\_armor', 't\_armor', 'ct\_money', 't\_money', 'ct\_helmets', 't\_helmets', 'ct\_defuse\_kits', 'ct\_players\_alive', 't\_players\_alive', 'ct\_weapon\_ak47', 't\_weapon\_ak47', 'ct\_weapon\_aug', 't\_weapon\_aug', 'ct\_weapon\_awp', 't\_weapon\_awp', 'ct\_weapon\_bizon', 't\_weapon\_bizon', 'ct\_weapon\_cz75auto', 't\_weapon\_cz75auto', 'ct\_weapon\_elite', 't\_weapon\_elite', 'ct\_weapon\_famas', 't\_weapon\_famas', 'ct\_weapon\_g3sg1', 't\_weapon\_g3sg1', 'ct\_weapon\_galilar', 't\_weapon\_galilar', 'ct\_weapon\_glock', 't\_weapon\_glock', 'ct\_weapon\_m249', 't\_weapon\_m249', 'ct\_weapon\_m4a1s', 't\_weapon\_m4a1s', 'ct\_weapon\_m4a4', 't\_weapon\_m4a4', 'ct\_weapon\_mac10', 't\_weapon\_mac10', 'ct\_weapon\_mag7', 't\_weapon\_mag7', 'ct\_weapon\_mp5sd', 't\_weapon\_mp5sd', 'ct\_weapon\_mp7', 't\_weapon\_mp7', 'ct\_weapon\_mp9', 't\_weapon\_mp9', 'ct\_weapon\_negev', 't\_weapon\_negev', 'ct\_weapon\_nova', 't\_weapon\_nova', 'ct\_weapon\_p90', 't\_weapon\_p90', 'ct\_weapon\_r8revolver', 't\_weapon\_r8revolver', 'ct\_weapon\_sawedoff', 't\_weapon\_sawedoff', 'ct\_weapon\_scar20', 't\_weapon\_scar20', 'ct\_weapon\_sg553', 't\_weapon\_sg553', 'ct\_weapon\_ssg08', 't\_weapon\_ssg08', 'ct\_weapon\_ump45', 't\_weapon\_ump45', 'ct\_weapon\_xm1014', 't\_weapon\_xm1014', 'ct\_weapon\_deagle', 't\_weapon\_deagle', 'ct\_weapon\_fiveseven', 't\_weapon\_fiveseven', 'ct\_weapon\_usps', 't\_weapon\_usps', 'ct\_weapon\_p250', 't\_weapon\_p250', 'ct\_weapon\_p2000', 't\_weapon\_p2000', 'ct\_weapon\_tec9', 't\_weapon\_tec9', 'ct\_grenade\_hegrenade', 't\_grenade\_hegrenade', 'ct\_grenade\_flashbang', 't\_grenade\_flashbang', 'ct\_grenade\_smokegrenade', 't\_grenade\_smokegrenade', 'ct\_grenade\_incendiarygrenade', 't\_grenade\_incendiarygrenade', 'ct\_grenade\_molotovgrenade', 't\_grenade\_molotovgrenade', 'ct\_grenade\_decoygrenade', 't\_grenade\_decoygrenade']

The dependent variable is the outcome variable and is measured in a study. Therefore, in this study, the dependent variable is “round\_winnder”.

Part 3: Data exploration (50 points)

1. List and describe three data exploration steps that you will likely need. (note: visualization is acceptable as three separate steps as long as they allow you to explore three different characteristics of the dataset) (30 points)
2. For each of the steps, describe how it will help you with pre-processing and analysis. (i.e., if my visualization indicates outliers, then I can remove them). (20 points)

Part 4: Pre-processing (50 points)

1. List and describe three processing steps that you think are involved in the data science pipeline. (30 points)
2. For each of the steps, what do you expect to achieve and how does it help your analysis? (20 points)

Part 5: Analysis (50 points)

1. In your own words, give an overview of the analysis you would need to perform to answer your data science question. (20 points)
2. Can you think of an algorithm learned in class that would be best to test your hypothesis? (30 points)
   1. If so, which one? Describe the algorithm
   2. If not, write down the algorithm

Part 6: Workflow (20 points)

1. Sketch the workflow for your data science problem, omitting the data exploration steps but including pre-processing and analysis. The workflow needs to include the data flow. The analysis step should be one component.

***IMPORTANT NOTES***

*Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences.  Please familiarize yourself with the discussion of plagiarism in SCampus in Section 11, Behavior Violating University Standards* [*https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions*](https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions/)*.  Other forms of academic dishonesty are equally unacceptable.  See additional information in SCampus and university policies on scientific misconduct,* [*http://policy.usc.edu/scientific-misconduct*](http://policy.usc.edu/scientific-misconduct/)*.*

*A number of USC’s schools provide support for students who need help with scholarly writing.  Check with your advisor or program staff to find out more.  Students whose primary language is not English should check with the American Language Institute* [*http://dornsife.usc.edu/ali*](http://dornsife.usc.edu/ali)*, which sponsors courses and workshops specifically for international graduate students.*

*For more information, see the class syllabus and the USC web site.*