Introduction:

The RMS Titanic was a British passenger liner that began its first voyage 100 years ago, becoming one of the most horrific maritime accidents after hitting an iceberg and sinking on its maiden voyage. The dataset that will be cleaned for this assignment contains 891 passengers on the Titanic. Though over 2000 passengers were aboard the vessel, data for only 891 exists. This highlights one of the problems with data science, namely the lack of data, especially historical data. Your job in this assignment is to clean this data and improve readability outlined in Part 3.

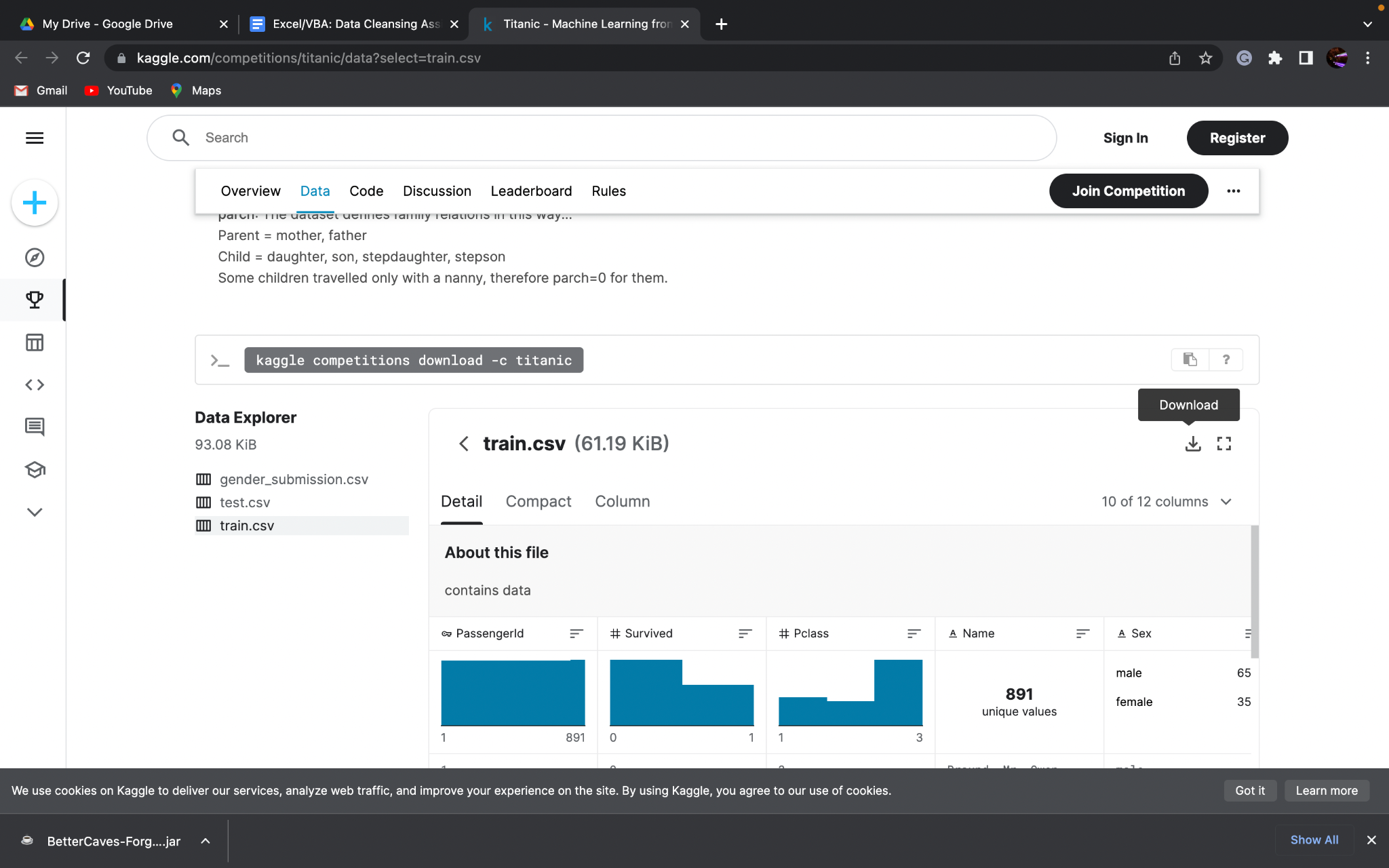
**Part 1: Downloading Excel**

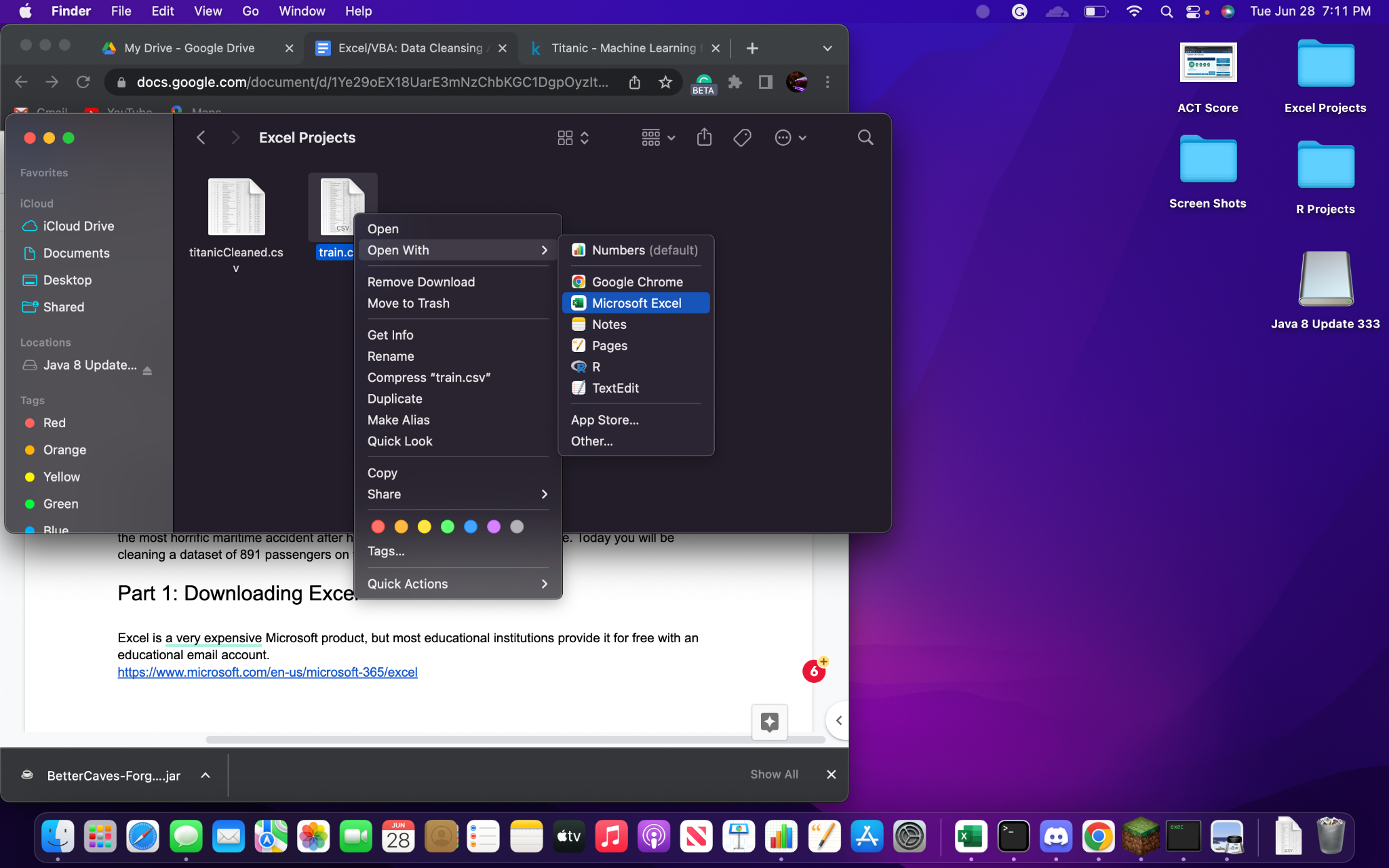
[Excel](https://www.microsoft.com/en-us/microsoft-365/excel) is a spreadsheet program similar to Google Sheets, but is much more powerful with the ability to create pivot tables, enhanced graphics creation, and macro programming using [Visual Basic for Applications (VBA)](https://docs.microsoft.com/en-us/office/vba/library-reference/concepts/getting-started-with-vba-in-office).

This assignment doesn’t involve pivot tables, graphics, or macro programming, instead, you will just be using base Excel functions to cleanse data.

**Part 2: Opening Excel and Downloading Data**

Once Excel has been downloaded, download the titanic dataset from [Kaggle](https://www.kaggle.com/competitions/titanic/data). The Titanic dataset you will be using is the train.csv file. Then open this file with Excel. Now you are ready to work.





After opening the file, you will see 12 variables with 891 observations of data. The [Kaggle](https://www.kaggle.com/competitions/titanic/data) page contains information about each variable and what they mean.

**Part 3: Cleansing the Data**

To cleanse this data, you must do the following:

1. Rename “Pclass” to “Socio-economic Status”, “SibSp” to “Siblings/Spouses”, “Parch” to “Parents/Children”, “Ticket” to “Ticket Number”
2. Check to make sure the data types in each column are as follows:
   1. “Fare” should be in Euros and rounded accordingly
   2. “Name” and “Sex” should be text
   3. All columns should contain appropriate data types
3. Deal with the blanks in the dataset
   1. For “Age” replace with the average of the existing ages
   2. For “Embarked” replace with most common value
   3. Since “Cabin” contains many NAs, keeping this variable in the dataset is generally not worthwhile. Not much can be done to make this data useable.
4. Change values and data types within columns
   1. “Survived” should be a boolean (“yes” or “no”)
   2. “Socio-economic Status” should be in the form of “upper”, “middle”, and “lower”, corresponding to 3, 2, and 1, respectively
   3. “Embarked” should have the full name of the embarkation point: “C” = “Cherbourg”, “Q” = “Queenstown”, “S” = “Southampton”

\*\*Note: Using [shortcuts](https://www.shortcutfoo.com/app/dojos/excel-win/cheatsheet) will help select data in columns much more efficiently. Microsoft published their own document of [Excel shortcuts](https://download.microsoft.com/download/5/4/c/54cd97c1-0213-46a7-b659-95f1c9145f42/Excel_shortcuts_cheat_sheet.pdf) that may be useful too.\*\*