# ROUND : DATA SCIENCE CASE-STUDY

## Introduction

This data set is a Beer data-set for your Data Science case-study round. You are expected to build a Machine Learning model which predicts the overall rating of the beer. ("review/overall" column in "train.csv" is your dependent variable.)

You are free to formulate this prediction problem either as a classification problem or regression problem.

## Inspiration

Here are a few questions which you may ask yourself and which may help you with this dataset:

1. How can you use "beer/name", "beer/style" and "review/text" as features to predict the overall rating of the beer ?
2. Are there any words that strongly predict the overall rating of the beer ?
3. How can you use other columns in train.csv to derive robust and effective features from them, which can help to predict the overall rating of the beer ?

## Expectations from this Task (expect in-depth discussions over your code, approach and methodologies in later rounds):

1. Data cleaning and Data preprocessing
2. Feature Engineering
3. Modelling using 1-2 ML models of your choice
4. At Least 2-3 Model Validation metrics

## Data fields

The train.csv contains the following columns:

- index - an identifier for the review
- beer/ABV - the alcohol by volume of the beer
- beer/beerId - a unique ID indicating the beer reviewed
- beer/brewerId - a unique ID indicating the brewery
- beer/name - name of the beer
- beer/style
- review/appearance - rating of the beer's appearance (1.0 to 5.0)
- review/aroma - rating of the beer's aroma (1.0 to 5.0)
- review/overall - rating of the beer overall (1.0 to 5.0)
- review/palate - rating of the beer's palate (1.0 to 5.0)
- review/taste - rating of the beer's taste (1.0 to 5.0)
- review/text - the text of the review
- review/timeStruct - a dict specifying when the review was submitted
- review/timeUnix
- user/ageInSeconds - age of the user in seconds
- user/birthdayRaw
- user/birthdayUnix
- user/gender - gender of the user (if specified)
- user/profileName - profile name of the user

## PREFERRED LANGUAGE:

Python or R