# DATA SCIENTIST SCREENING TEST QUESTIONS

As part of the recruitment process of a data scientist, the first level of screening involves a series of questions which revolves around the life of a data scientist, testing you from the fundamentals to the core elements of machine learning.

## GENERAL QUESTIONS

This section involves basic programming and general theoretical tasks.

1. Write a simple program to implement selection sort.
2. Given a function rand5() that returns a random int between 0 and 5, implement rand7()
3. Write a function that sorts a stack (bonus: sort the stack in place without extra memory)
4. Implement a linked list (with insert and delete functions)
5. Give examples of data cleaning techniques you have used in the past.
6. Could you please provide details about precision and recall in a machine learning context with an example?
7. If the CEO of a company asks you what model ‘Lift’ is, how would you explain it to him?

## DATA EXPLORATION CHALLENGE

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset, which is exactly what this task demands.

The task is to download the dataset and do an exploratory analysis by using either R or Python. The dataset can be downloaded from [**Beer Reviews**](https://jadatasciencetest.blob.core.windows.net/datasciencetest/beer_reviews_data.csv)

Some of the questions you will need to answer based on the data set are:

1. Which brewery produces the strongest beers by ABV%?

2. If you had to pick 3 beers to recommend using only this data, which would you pick?

3. Which of the factors (aroma, taste, appearance, palate) are most important in determining the overall quality of a beer?

4. Lastly, if I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

Bonus points if you can come up with your own set of questions and relevant insights.

## DATA SCRAPING AND TEXT ANALYSIS

This task is to test your mettle in the areas of using APIs and then subsequently performing text analysis. You are required to scrape data from **Twitter API** for the following two airlines and store the data in a flat file.

* Air India
* Singapore Airlines

Further you are required to perform the following text analysis on the scraped data and answer the following questions (Applicable for both the airlines):

1. Are the customers satisfied with the services offered by the airlines?
2. Can you extract the most relevant reasons as stated by the tweets related to why the customers might be satisfied or dissatisfied with the airlines?

For each of the questions above, please attach relevant codes written in R/Python 2/3, appropriate visualizations along with a short write-up about the approach taken.

*(Twitter API Link :* [*https://developer.twitter.com/en/docs*](https://developer.twitter.com/en/docs)*)*

## TIME SERIES ANALYSIS & FORECASTING

The given dataset consists of cheese production volumes for three different types of cheese from the year 1995 to 2013. The data is clean but due to data collection issues has missing data for a few months. Your task is to perform a time series analysis on the historical data and forecast the cheese production volumes for the next year one. You can write your codes in Python 2/3 or R.

Also document the steps you took during data cleaning, imputation if any, analysis and forecasting and why you think the method you used is good for this scenario over other methods you might be aware of.

Dataset (Use the files provided if you cannot access the dataset below)



## CODING CHALLENGE – FEATURE ENGINEERING & MACHINE LEARNING

The following dataset (csv) consists of credit card customers belonging to a financial institution in Taiwan. It consists of payment, history, credit and demographic information tagged to whether the card defaulted or not previously. The task is to build a prediction model around this data to predict the probability of default. From the perspective of risk management, the probability is considered more valuable than the binary classification itself. *You are free to use either Python 2/3 or R for coding.*

You will be tested on:

1. Feature Engineering –
   1. What features would you derive?
   2. How you evaluate those features?
   3. Techniques used.
2. Machine Learning Models –
   1. Different algorithms and approach
3. Optimizing the Models –
   1. What are the different approaches used?

Evaluation

You will be evaluated based on your results on the test data.

Please write theoretical answers where you see fit in an accompanying document.

Attached Datasets

For training - 

For testing - 