# Scrum Cycle: Working with Python or R

# CASE STUDY

BankN is keen on improving their customer targeting and develop new business lines in the face of reduced revenues brought on by COVID-19 restrictions on businesses. Given the downturn in transactions and the reduced number of persons taking loans and investing, they need your help. They have been collecting extended data about their customers prior to the pandemic and would like to explore the data to assist them in some key decision-making activities towards increased revenues.

You have been asked to work in a team to explore and prepare the data for more complex analysis, hopefully, leading to meaningful actionable insights. However, you are also expected to assist the Bank to discover some preliminary insights and answer some key questions. The dataset provides several observations each with features as described in table I. This data was extracted from the bank's database at the end of 2019, this to ensure the pandemic period was excluded.

**Table I: Data Dictionary**

| Column Name | Description |
|---|---|
| RefNum | A unique identifier for each customer |
| age | numeric age of the customer |
| job | Type of Job ("admin", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services") |
| marital | Marital status ("married","divorced","single"; note: "divorced" means divorced or widowed) |
| education | ("secondary","primary","tertiary") |
| housing | Does this person have a mortgage |
| loan | Does this person have a personal loan |
| day | Last deposit day of the month (numeric) |
| month | Last deposit month of year |
| duration | Time spent on call when Last contact was done with customer, in seconds (numeric) |
| deposit | Amount deposited by customer in the last transaction |
| balance | Average yearly balance in Jamaican dollars |

The dataset to be used by each team is a different subset assigned as follows:

| Team | Assigned Dataset |
|---|---|
| Agile Agents | 2021-Set01-AgileAgents |
| Code Titans | 2021-Set02 - CodeTitans |
| Quality Assured | 2021-Set03-QualityAssured |

**REQUIRED:**

1. Load the dataset and explore the contents, summarizing your findings of the data quality in relation to standard measures (including, mean, median, standard deviation, missing values, outliers, etc.). Review the data dictionary attached to get the context of each measurement.
    a. Describe the data in terms of the types and measurements and what each measurement represents (purpose of features, range of values, min, max, etc.)
    b. Fix missing values, noise and outliers as necessary
2. Convert the `duration` from the current value in `seconds` to `minutes`.
3. Apply one of the binning methods discussed to discretize `age`. Justify your selection and apply the changes to make `age` a categorical variable (factor in R).
4. Apply one of the `normalization` methods discussed to any feature in the dataset. Justify your decision.
5. Construct a new column (variable) called `last_deposit` which records the number of days since the client made a deposit. Use the month and day columns and calculate this value with the current date.
6. Plot time-series graph(s) to show comparison between deposits made over 2 comparative periods within the data (month/quarter)

7. **<u>Use the data to answer to the following:</u>**
    a. Which categories of jobs have the highest and lowest average deposits?
    b. What is the dominant educational level of each category of job?
    c. What percentage of persons have both **mortgage** and **personal loan**?
    d. Which age-group has the highest average balance?
    e. Is there a correlation between **deposit** and **balance**? Discuss the findings. Explain the implications.
    f. What percentage of customers who are married also are in overdraft?
    g. If the Bank wanted to launch a new loan product to target persons who may be considered to have the least credit risk, propose a metric to determine these persons. Use at least 3 features to justify/select your target profile. For example: *RefNum < X AND Married AND education=="Tertiary".*
        i. You must be prepared to explain why you made the decision on the variables used for determining the best person to target. This may be helped by investigating what banks generally use.
        ii. You should extract the subset of data that meets the profile (this could be considered a lead list to be used by customer service to contact these customers). Consider that you will need to provide the bank with the RefNum for the customers they should be targeting.

NOTE: For presentation purposes, be prepared to show results of all steps you carried out including visuals of the results.