Candidate’s Name:

Please note the following instructions:

1. Two datasets **spenddata.csv** and **testdata.csv** are provided for Question 1**.** You are free to decide how best to use them.
2. Refer to **mock survey data 3.xls** for Question 2.
3. All source codes must be provided with comments along with instructions on how to execute the code.
4. R is preferred, but you are free to consider other tools. Regardless of the tool you use, source code / workings must be provided and clearly documented
5. You do not need to augment this dataset with any external data. However, if you choose to do so, you must document the reasons.

**Question 1**

Some of the respondents have been tagged as belonging to group 1 – 6. However, due to a data calculation issue, some of the respondents have had their groups (**pov6**) missing.

Build a model that will classify these respondents back into one of the 6 groups.

1. Please explain the choice of metric / evaluation criterion used
2. What are the assumptions you made when building this model?
3. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.
4. Please explain under what conditions will the model you choose **be not appropriate**
5. How confident are you of the model’s robustness and how would you explain the model’s performance?
6. Why is your model performing well / not well?
7. Was any feature engineering required? If yes, what were they. If no, why?

Please include your supporting scripts/output from R or an alternative tool of your choice.

**Question 2**

Using the mock survey data (mock survey data 3.xlsx), answer the following questions. Note:

* a5 - Nationality
* totfnb\_$ - FnB spend
* totexp\_$ - Total Expenditure

1. Design an efficient schema to allow the users to query the data efficiently. You will need to provide an ERD of the proposed schema. The design should take into consideration the following:
   1. V Hotel Lavender and V Hotel Bencoolen was merged to be part of V Hotel Group in the later part of the year
   2. There is a need to still analyse data based on the hotels before the merger
2. Based on the schema, construct the necessary SQL(s) to answer the following business questions:
   1. What is the total expenditure trend of visitors who stayed in V Hotel by month?
   2. Which nationality has the highest FnB spend for each of the following:
      1. V Hotel Lavender
      2. V Hotel Bencoolen
      3. V Hotel Group
   3. Is there a correlation between travel companions and choice of hotel?
   4. In addition to your findings from (a - c), what other insights can be derived from analysing business and leisure visitors and how will your insights help STB attract visitors to spend in Singapore. **Prepare a short PowerPoint presentation to share this**. *Hints:*
      1. *The type of visitors can be found through purpose of visit)?*
      2. *You are encouraged to explore the data beyond the questions asked in a – c*
      3. *Use appropriate charts / visuals to communicate your story*
      4. *Note the chart-ink ratio*
3. Due to data collection issues, there are many records with outliers that need to be removed. Suggest a method to identify outliers in the dataset and explain the steps you have taken to arrive at that method and why you have chosen that method. Please include any codes if needed.