|  |  |
| --- | --- |
| Cairo University  Faculty of Computers and Information  Data science and big data analytics  Use Case –Workflow 2 |  |

**Data processing phase (4 marks)**

1. In RStudio , read the table you have created in Workflow1 (padat table) (0.25 mark)
2. In R, turn some of the codes into factors for those variables highlighted in Workflow1 step 5, what is the importance of factorization? Plot some of the variables you will create. (0.75 mark)
3. How many people were accepted in the Adams county (0.5 mark)
4. From the query in Workflow 1, we inserted the income and other variables as VARCHAR , where it should be numeric, convert the needed variables to numeric. Make sure that they are numeric using a function (1 mark)
5. Remove records without income info. ( 0.5 mark)
6. Remove rows with nulls (NA) in Tract\_To\_MSAMD\_Income\_pct, Minority\_Population\_pct, Tract\_To\_MSAMD\_Income\_pct (1 mark)

**Data Analysis phase (6 marks)**

1. Visualize at least 5 variables and elaborate the results for each plot (2.5 marks)
2. Is there a relation between the applicant income and the loan amount, elaborate? (0.5 mark)
3. Check whether the loan amount has any odd, multi-modal distribution. This may suggest to us that we might want to build separate models for the different loan purposes.(1 marks)
4. What is the loan purpose type that we need drop to make the experiment cleaner? (0.5 mark)
5. Look for spikes in the outputs. You may need to develop one model with all the data below that spike and develop a separate model with the data beyond that spike ( You can check this on the loan amount ink variable) (1.5 marks)