ECOM5005 – Business Analytics RStudio Report

Group 0: Student Name 1 (ID), Student Name 2 (ID), Student Name 3 (ID)

Due on xx October 2025 at 23:59 AWST

Main Instructions (Please read carefully)

- **Group work**: This is a group assessment, and each group must consist of **three (3) students**. Students are responsible for forming their own groups.
- **Group sign-up form submission**: Each group must complete the **Group Sign-Up Form** and email it to the unit address: **ecom5005@curtin.edu.au** by the end of Week 5. The form is available inside the Assessment folder on Blackboard.
- **Use of ChatGPT**: Students are **expected to use ChatGPT** to assist with their analysis and reporting throughout this assessment.
  - The **free version of ChatGPT** offers sufficient access to daily prompts and file attachment options, enabling efficient support for tasks such as code explanation, data interpretation, and improving written communication.
  - While the use of ChatGPT is encouraged to enhance the learning experience, students are reminded to critically assess and validate any AI-generated content used in their report.
- **Support through tutorials and AI tools**: All tasks in this assessment have been demonstrated to some extent during tutorials. Students may use AI tools like ChatGPT or others to enhance their data visualisation or interpretation, although doing so is not a requirement to achieve a Distinction or above.
- **Submission Requirements**: Each group must submit **one final PDF file** and **Student Gen-AI Disclosure Form** (one per student):
  - A **PDF** version of your completed **RMarkdown report** (exported from the starter file).
  - Submit your file via the **Turnitin submission point**, which will be available in your **Group folder** or the **Assessment folder** on Blackboard.
  - **Only PDF format is accepted**. Submissions in other formats (e.g., .Rmd, .docx) will not be marked. **Only one student per group needs to upload the file**.
  - Each student is required to complete and submit the **Gen-AI Disclosure Form**, ensuring that the **"Agree"** option is selected at the end.
  - Since most groups will consist of **3 members**, each group must upload a total of **4 files**: one RMarkdown PDF report and three individual Gen-AI Disclosure Forms.
  - A reminder and submission instructions will also be posted on Blackboard closer to the due date.

RStudio Instructions

- **Start from the provided RMarkdown file:**

- You must start your assessment from the RMarkdown file named **ECOM5005-Assessment-Student-Depression-Starter**. This file contains the required structure and guidance to complete the task. Do not create your own file from scratch. Failure to use the starter file may result in incomplete or improperly formatted submissions.

1. Except for your Group Number, Student Names and IDs, **DO NOT** change anything in the YAML above.

2. You have to install `tinytex` if not installed already.

- First Type: install.packages("tinytex") in the console.
- Second, Type: tinytex::install_tinytex().
- Restart R after the installation.

3. Install any other packages you may need for the assessment using the "Packages" tab in the output panel or the "install.packages()" function in the console.

4. Make sure your knitted PDF output is well organised as much as possible. Try and break to the next line, any code or text in the code chunks that exceed the code chunk margins.

5. Your responses and interpretations should be in the text section. Remember you can still implement a simple code in the text chunks – Hints have been given in this regard where necessary. **Stick to the Word Limits** to the responses to avoid losing points. All your main codes should be in the code chunks. Follow the code chunks and the sub-headings with instructions carefully to answer the questions. **AVOID** as much as possible creating additional sub-headings in the code chunks.

6. Be mindful AI tools may give wrong answers. It is your responsibility to verify the output for correctness. You are to insert 3 samples of the Prompts you used in the AI tool and the responses from the AI tool. The relevant sections requiring these samples are clearly indicated.

7. It is a good practice to occasionally knit your file to make sure the PDF is knitted successfully.

8. Clarity of the report will attract points.

Assessment Goal

The goal of this group assessment is to develop a predictive model given the **student_depression_dataset.csv** dataset depending on your group. Students will explore the dataset, select appropriate predictors, build a statistical model in R, and present their findings in a structured report which is the knitted pdf file.

The assessment emphasises data preprocessing, thoughtful predictor selection, model building, and clear communication of results. The assessment also introduces students to how AI tools can be incorporated into Business Analytics.

The structure of this R Markdown document has text sections and code chuncks. Provide any text responses or interpretations in the text sections.

1 Exploratory Data Analysis (EDA)

1.1 Import the student_depression_dataset.csv dataset assigned to your group and assign it to the object "depress" and take a look at the dataset. (0.5 Point)

1.2 1) How many variables are in your dataset? how many of them are categorical or character variables? and 2) How many missing data points are in your dataset? (0.5 points)

(Hint: You can use in-text r code `r` to insert a code in the text section)

(Hint: you use ncol, sum and sapply functions) 1st Response here (Number of variables and categorical/character variables):

2nd Response here (Number of missing observations):

1.3 Clean your dataset of all missing data points if any and assign this to a new object name: "depress_clean". Then verify that there are no missing data points in the cleaned dataset. (0.5 Points)

Verify the total number of missing observations in the cleaned dataset.

(Hint: You can use in-text r code `r` to insert a code in the text section)

Response here:

1.4 Check the levels of each of the character variables. Use the AI prompt and provide the screenshot after the code chunk (0.75 Point).

1.4.1 Insert a screenshot of the Prompt from the AI model for Levels check below.

Hint: use `![Caption](Directory){ width=50% }`. Insert the caption of the figure and the directory to the image file accordingly. Feel free to change the width size to fit nicely in your knitted PDF file.

1.4.2 Do you think the Depression variable should be an integer type date as it is in the dataset? (0.25 Points)

Respond here (Max. 20 Words):

1.5 Since we want to work through the variables in the dataset, carefully examine the dataset and create factor variables where necessary. Verify afterwards that the factor variables were created succesfully. (0.5 Point)

1.6 Now use the following short variable names for the variables in the dataset: (0.5 Points)

Have.you.ever.had.suicidal.thoughts.. → Suicide

Family.History.of.Mental.Illness → Mental

Sleep.Duration → Sleep

Dietary.Habits → Diet

Depression → Depress

Profession → Prof

Academic.Pressure → Acad

Work.Pressure → Work

Study.Satisfaction → Study

Job.Satisfaction → Job

Work.Study.Hours → WorkHrs

Financial.Stress → FinStress

1.7 We may not need to use id. Also, it seems Work Pressure and Job Satisfaction have only zeros. Drop id, Work and Job from the dataset. Also remove all observations with CGPA of zero (may influence negatively the data analysis). (0.5 Points)

2 Data Visualisation

Our first goal will be to predict **CGPA**. before that let's do some data visualisation considering this variable and the predictors.

2.1 First, plot the distribution of CGPA by depression status of the person. (0.5 Point)

2.1.1 Comment on the distribution of CGPA based on the depression status. (0.5 Points)

Response here (Max. 30 words):

2.2 Take a graphical look at the pairwise correlation between all numeric variables in the dataset. Use the AI model and provide a screenshot of the prompt and response after the code chunk. (0.75 Points)

2.2.1 Insert a screenshot of the Prompt from the AI model for Correlation matrix below.

Hint: use `![Caption](Directory){ width=60% }`. Insert the caption of the figure and the directory to the image file accordingly. Feel free to change the width size to fit nicely in your knitted PDF file.

2.2.2 (1) Give a general comment about the correlations. (2) Also comment on the correlations of the predictors with the dependent variable. Which of them is likely to be a good predictor (0.5 treshold)? (3) Concerning the predictors, is there potential for multicollinearity (0.7 treshold)? Explain. (0.75 Points)

Response here:

(1) General comments (Max. of 20 words):

(2) Predictors and outcome variable (Max. of 50 words):

(3) Is there potential of multicollinearity? Explain (Max. of 20 words):

2.3 Distribution of Depression Status

For our second goal, we will be predicting **depression**. Hence, let's explore this dependent variable.

2.3.1 Table and Bar chart: Count and Proportion of Depression Status

Provide a table that shows the **count** and **proportion** of **depression status**. (1.5 Points)

2.3.2 Now plot a bar chart to show the count of depressed and undepressed individuals. (1 Point)

3 Multiple linear regression model of CGPA on all other variables. The regression model is speficied below.

$$
\begin{aligned}
CGPA = {} & \beta_0 + \beta_1\text{Gender} + \beta_2\text{Age} + \beta_3\text{City} + \beta_4\text{Prof} + \beta_5\text{Acad} \\
& + \beta_6\text{Study} + \beta_7\text{Sleep} + \beta_8\text{Diet} + \beta_9\text{Degree} \\
& + \beta_{10}\text{Suicide} + \beta_{11}\text{WorkHrs} + \beta_{12}\text{FinStress} + \beta_{13}\text{Mental} \\
& + \beta_{14}\text{Depress} + \epsilon
\end{aligned}
$$

3.1 Now estimate the multiple linear regression model and assign it to an object "CGPA_model" using the new dataframe "depress_clean". Take a look at the regression output.(1 Points)

3.2 Provide a brief interpretation of the regression summary following the questions below. (1.5 Point)

(1) Are the predictors jointly significant in predicting **CGPA**?

Response here (Max. 20 words):

(2) What percentage variation in **CGPA** is explained by the predictors?

Response here (Max. 20 words):

(3) Interpret the coefficient of the predictor **Acad**

Response here (Max 20 words):

3.3 It seems the model is not parsimonuos. We will perform and stepwise regression in this section.

3.3.1 Let's perform backward stepwise model selection using the function `step()` based on the depress_clean dataset. (1 Point)

3.3.2 How many predictors are chosen using the backward stepwise regression? Identify these predictors. (0.5 Points)

Response here (Max. 20 words):

3.3.3 Estimate the selected model and assign to the object "CGPA_model2". (0.5 Points)

3.3.3.1 Comment on the performance of the selected model using the coefficient of determination. (0.5 Points) Response here (Max. 20 words):

3.3.3.2 Interpret the coefficeint of Gender. (0.5 Points) Response here (Max. 20 words):

3.4 Let's check the performance of our two models using the root mean squared error (RMSE).

3.4.1 Do an in-sample prediction of CGPA for both models and estimate the RMSE for both models and comment on the results. (2.5 Points)

3.4.1.1 Comment on the RMSE values. (0.5 Points) Response here: (Max. 30 words):

4 Logistic regression. This time we want to specify depression as a function of all the predictors.

4.1 First let's split the updated data into train and test. Perform a 70/30 split of the updated dataset into train/test splits. It is important to ensure that both datasets should contain similar proportions of depressed and non-depressed individuals and similar proportions of the profession. Hence, create a new variable which is an interaction between Depress and Prof and use this new variable to ensure the similar proportions. REMEMBER to drop the new variable from the train and test dataset. Also, in order to ensure replicability, set seed to 123. Print the proportion of depression in the train and test datasets. (4 Points)

4.2 Now, in order to ensure a parsimonuos model, let's perform a forward stepwise model selection using the function `step()` based on the train dataset. (1.5 Points)

4.2.1 How many predictors are chosen using the forward stepwise regression? List them. (0.5 Points)

Response here (Max. 20 words):

4.2.2 Using the selected model from the forward stepwise regression, fit the final selected logistics regression model. (0.5 Points)

4.2.3 Coefficient Interpretations. Let's interpret four coefficients. Round your estimates to 2 decimal places (2 Points). Hint: Use the `r` to call for the coefficient and estimate the effect.

(1) If **Acad** increases by 1 unit:

Response here (Max. 20 words):

(2) If **Age** increases by 1 year:

Response here (Max. 20 words):

(3) If the student **Sleep'Less than 5 hours'**:

Response here (Max. 20 words):

(4) If **WorkHrs** Reduce by 3 hour:

Response here (Max. 20 words):

4.3 Evaluate the model performance using the test dataset by using threshold 0.5 based on the following metrics: (1) Accuracy; (2) Sensitivity; and (3) Specificity. (1.5 Points)

4.3.1 Insert a screenshot of the Prompt from the AI model for model performance below.

Hint: use `![Caption](Directory){ width=60% }`. Insert the caption of the figure and the directory to the image file accordingly. Feel free to change the width size to fit nicely in your knitted PDF file.

4.4 Interpret the three evaluation metrics. (1.5 Points)

Response here:

Accuracy (Max. 20 words):

Sensitivity (Max. 20 words):

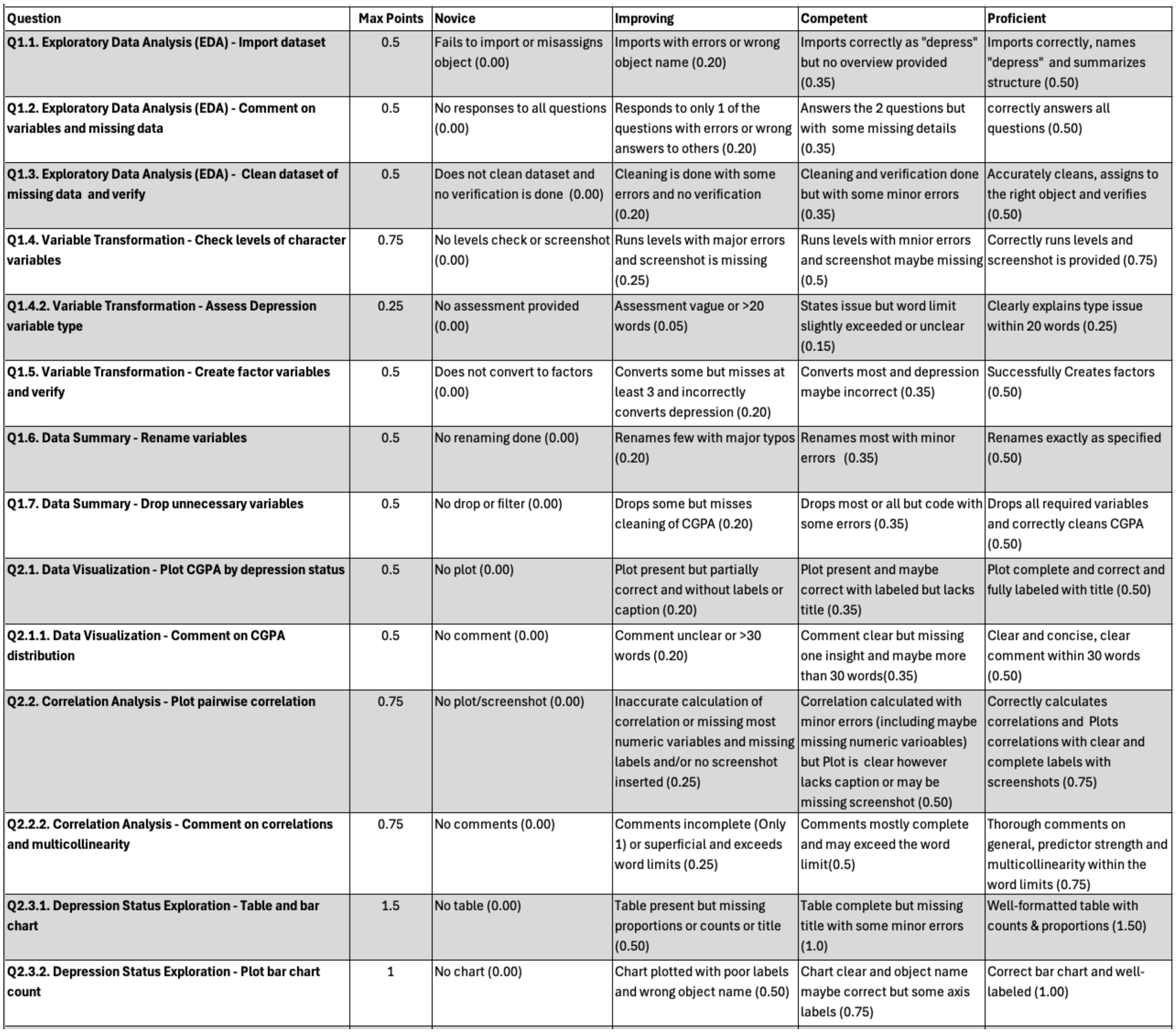
Specificity (Max. 20 words):

4.5 Plot the ROC curve. (0.5 Points)

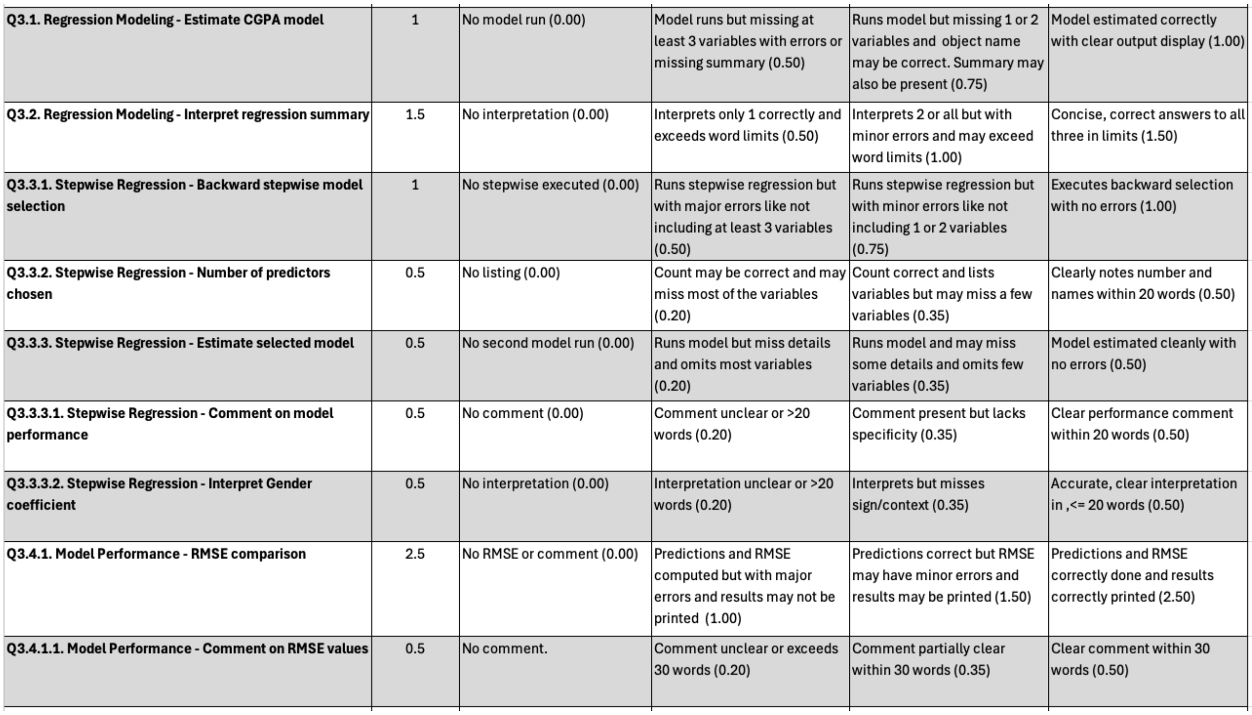
4.5.1 Interpret the AUC value (0.5 Points):

Response here (Max. 50 words):

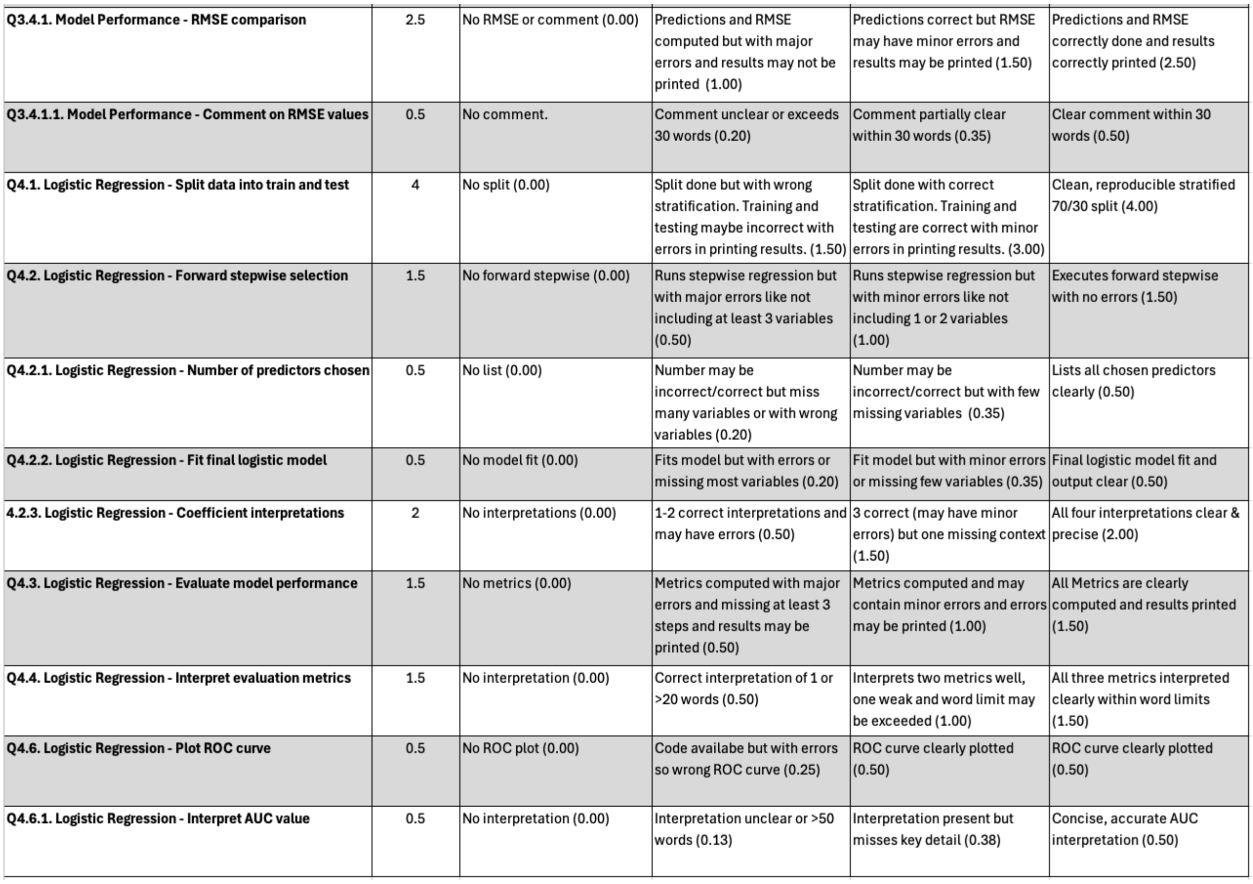
5 Marking Rubric



Question	Max Points	Novice	Improving	Competent	Proficient
Q1.1. Exploratory Data Analysis (EDA) - Import dataset	0.5	Fails to import or misassigns object (0.00)	Imports with errors or wrong object name (0.20)	Imports correctly as "depress" but no overview provided (0.35)	Imports correctly, names "depress" and summarizes structure (0.50)
Q1.2. Exploratory Data Analysis (EDA) - Comment on variables and missing data	0.5	No responses to all questions (0.00)	Responds to only 1 of the questions with errors or wrong answers to others (0.20)	Answers the 2 questions but with some missing details (0.35)	correctly answers all questions (0.50)
Q1.3. Exploratory Data Analysis (EDA) - Clean dataset of missing data and verify	0.5	Does not clean dataset and no verification is done (0.00)	Cleaning is done with some errors and no verification (0.20)	Cleaning and verification done but with some minor errors (0.35)	Accurately cleans, assigns to the right object and verifies (0.50)
Q1.4. Variable Transformation - Check levels of character variables	0.75	No levels check or screenshot (0.00)	Runs levels with major errors and screenshot is missing (0.25)	Runs levels with mnior errors and screenshot maybe missing (0.5)	Correctly runs levels and screenshot is provided (0.75)
Q1.4.2. Variable Transformation - Assess Depression variable type	0.25	No assessment provided (0.00)	Assessment vague or >20 words (0.05)	States issue but word limit slightly exceeded or unclear (0.15)	Clearly explains type issue within 20 words (0.25)
Q1.5. Variable Transformation - Create factor variables and verify	0.5	Does not convert to factors (0.00)	Converts some but misses at least 3 and incorrectly converts depression (0.20)	Converts most and depression maybe incorrect (0.35)	Successfully Creates factors (0.50)
Q1.6. Data Summary - Rename variables	0.5	No renaming done (0.00)	Renames few with major typos (0.20)	Renames most with minor errors (0.35)	Renames exactly as specified (0.50)
Q1.7. Data Summary - Drop unnecessary variables	0.5	No drop or filter (0.00)	Drops some but misses cleaning of CGPA (0.20)	Drops most or all but code with some errors (0.35)	Drops all required variables and correctly cleans CGPA (0.50)
Q2.1. Data Visualization - Plot CGPA by depression status	0.5	No plot (0.00)	Plot present but partially correct and without labels or caption (0.20)	Plot present and maybe correct with labeled but lacks title (0.35)	Plot complete and correct and fully labeled with title (0.50)
Q2.1.1. Data Visualization - Comment on CGPA distribution	0.5	No comment (0.00)	Comment unclear or >30 words (0.20)	Comment clear but missing one insight and maybe more than 30 words(0.35)	Clear and concise, clear comment within 30 words (0.50)
Q2.2. Correlation Analysis - Plot pairwise correlation	0.75	No plot/screenshot (0.00)	Inaccurate calculation of correlation or missing most numeric variables and missing labels and/or no screenshot inserted (0.25)	Correlation calculated with minor errors (including maybe missing numeric varioables) but Plot is clear however lacks caption or may be missing screenshot (0.50)	Correctly calculates correlations and Plots correlations with clear and complete labels with screenshots (0.75)
Q2.2.2. Correlation Analysis - Comment on correlations and multicollinearity	0.75	No comments (0.00)	Comments incomplete (Only 1) or superficial and exceeds word limits (0.25)	Comments mostly complete and may exceed the word limit(0.5)	Thorough comments on general, predictor strength and multicollinearity within the word limits (0.75)
Q2.3.1. Depression Status Exploration - Table and bar chart	1.5	No table (0.00)	Table present but missing proportions or counts or title (0.50)	Table complete but missing title with some minor errors (1.0)	Well-formatted table with counts & proportions (1.50)
Q2.3.2. Depression Status Exploration - Plot bar chart count	1	No chart (0.00)	Chart plotted with poor labels and wrong object name (0.50)	Chart clear and object name maybe correct but some axis labels (0.75)	Correct bar chart and well-labeled (1.00)



Q3.1. Regression Modeling - Estimate CGPA model	1	No model run (0.00)	Model runs but missing at least 3 variables with errors or missing summary (0.50)	Runs model but missing 1 or 2 variables and object name may be correct. Summary may also be present (0.75)	Model estimated correctly with clear output display (1.00)
Q3.2. Regression Modeling - Interpret regression summary	1.5	No interpretation (0.00)	Interprets only 1 correctly and exceeds word limits (0.50)	Interprets 2 or all but with minor errors and may exceed word limits (1.00)	Concise, correct answers to all three in limits (1.50)
Q3.3.1. Stepwise Regression - Backward stepwise model selection	1	No stepwise executed (0.00)	Runs stepwise regression but with major errors like not including at least 3 variables (0.50)	Runs stepwise regression but with minor errors like not including 1 or 2 variables (0.75)	Executes backward selection with no errors (1.00)
Q3.3.2. Stepwise Regression - Number of predictors chosen	0.5	No listing (0.00)	Count may be correct and may miss most of the variables (0.20)	Count correct and lists variables but may miss a few variables (0.35)	Clearly notes number and names within 20 words (0.50)
Q3.3.3. Stepwise Regression - Estimate selected model	0.5	No second model run (0.00)	Runs model but miss details and omits most variables (0.20)	Runs model and may miss some details and omits few variables (0.35)	Model estimated cleanly with no errors (0.50)
Q3.3.3.1. Stepwise Regression - Comment on model performance	0.5	No comment (0.00)	Comment unclear or >20 words (0.20)	Comment present but lacks specificity (0.35)	Clear performance comment within 20 words (0.50)
Q3.3.3.2. Stepwise Regression - Interpret Gender coefficient	0.5	No interpretation (0.00)	Interpretation unclear or >20 words (0.20)	Interprets but misses sign/context (0.35)	Accurate, clear interpretation in ,<= 20 words (0.50)
Q3.4.1. Model Performance - RMSE comparison	2.5	No RMSE or comment (0.00)	Predictions and RMSE computed but with major errors and results may not be printed (1.00)	Predictions correct but RMSE may have minor errors and results may be printed (1.50)	Predictions and RMSE correctly done and results correctly printed (2.50)
Q3.4.1.1. Model Performance - Comment on RMSE values	0.5	No comment.	Comment unclear or exceeds 30 words (0.20)	Comment partially clear within 30 words (0.35)	Clear comment within 30 words (0.50)



Q3.4.1. Model Performance - RMSE comparison	2.5	No RMSE or comment (0.00)	Predictions and RMSE computed but with major errors and results may not be printed (1.00)	Predictions correct but RMSE may have minor errors and results may be printed (1.50)	Predictions and RMSE correctly done and results correctly printed (2.50)
Q3.4.1.1. Model Performance - Comment on RMSE values	0.5	No comment.	Comment unclear or exceeds 30 words (0.20)	Comment partially clear within 30 words (0.35)	Clear comment within 30 words (0.50)
Q4.1. Logistic Regression - Split data into train and test	4	No split (0.00)	Split done but with wrong stratification. Training and testing maybe incorrect with errors in printing results. (1.50)	Split done with correct stratification. Training and testing are correct with minor errors in printing results. (3.00)	Clean, reproducible stratified 70/30 split (4.00)
Q4.2. Logistic Regression - Forward stepwise selection	1.5	No forward stepwise (0.00)	Runs stepwise regression but with major errors like not including at least 3 variables (0.50)	Runs stepwise regression but with minor errors like not including 1 or 2 variables (1.00)	Executes forward stepwise with no errors (1.50)
Q4.2.1. Logistic Regression - Number of predictors chosen	0.5	No list (0.00)	Number may be incorrect/correct but miss many variables or with wrong variables (0.20)	Number may be incorrect/correct but with few missing variables (0.35)	Lists all chosen predictors clearly (0.50)
Q4.2.2. Logistic Regression - Fit final logistic model	0.5	No model fit (0.00)	Fits model but with errors or missing most variables (0.20)	Fit model but with minor errors or missing few variables (0.35)	Final logistic model fit and output clear (0.50)
4.2.3. Logistic Regression - Coefficient interpretations	2	No interpretations (0.00)	1-2 correct interpretations and may have errors (0.50)	3 correct (may have minor errors) but one missing context (1.50)	All four interpretations clear & precise (2.00)
Q4.3. Logistic Regression - Evaluate model performance	1.5	No metrics (0.00)	Metrics computed with major errors and missing at least 3 steps and results may be printed (0.50)	Metrics computed and may contain minor errors and errors may be printed (1.00)	All Metrics are clearly computed and results printed (1.50)
Q4.4. Logistic Regression - Interpret evaluation metrics	1.5	No interpretation (0.00)	Correct interpretation of 1 or >20 words (0.50)	Interprets two metrics well, one weak and word limit may be exceeded (1.00)	All three metrics interpreted clearly within word limits (1.50)
Q4.6. Logistic Regression - Plot ROC curve	0.5	No ROC plot (0.00)	Code availabe but with errors so wrong ROC curve (0.25)	ROC curve clearly plotted (0.50)	ROC curve clearly plotted (0.50)
Q4.6.1. Logistic Regression - Interpret AUC value	0.5	No interpretation (0.00)	Interpretation unclear or >50 words (0.13)	Interpretation present but misses key detail (0.38)	Concise, accurate AUC interpretation (0.50)

The End