# MODULE 1 UNIT 2

## Activity submission

Learning outcomes:

LO3: Recognise the different types of data.

LO4: Interpret given data through suitable visualisations.

LO5: Analyse data in R in preparation for machine learning applications.

|  |
| --- |
| **Plagiarism declaration:** |
| **1. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own.**  **2. This assignment is my own work.**  **3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.**  **4. I acknowledge that copying someone else’s assignment (or part of it) is wrong, and declare that my assignments are my own work.** |

### Name:

#### 1. Instructions and guidelines (Read carefully)

##### Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M1 U2 Activity Submission** – **e.g. Lilly Smith M1 U2 Activity Submission.** **NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*
2. Write all your answers in this document. There is an instruction that says, “Start writing here” under each question. Please type your answer there.
3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.
4. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

**PLEASE NOTE:** **Plagiarism cases will be investigated in line with the Terms and Conditions for Students.**

IMPORTANT NOTICE: Please ensure that you have checked your course calendar for the due date for this assignment.

##### Guidelines

1. There are five pages and one question in this assignment.
2. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully but concisely and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).
3. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**
4. At the end of your assignment, please provide feedback on areas where you require further assistance or would like the Assessor to expand on.

#### 2. Mark allocation

The question counts 18 marks. However, you will only receive a final percentage mark and will not be given individual marks for the sections of the question. Use the grading rubric to see how marks will be allocated.

#### 3. Contextual information

You are a senior financial analyst at a financial institution, FundU, which offers credit solutions to customers. You have collected data regarding customers who have defaulted on their credit payments over the past year.

After you have successfully imported data into R, detected issues with the data, cleaned the data, and created the required plots, you are ready to analyse the generated plots to inform a decision.

##### Question 1

Having generated the different plots in the IDE notebook, analyse these plots and answer the following questions:

* 1. With specific reference to the type of data in the credit data set, are these plots being used properly? Substantiate the appropriate use of the plots generated to visualise the data.
  2. Compile two general personas: one of the typical individuals you will award a loan to, and another of the typical individuals you will not award a loan to. To answer this question, analyse the plots generated in the IDE Activity (Assessment) to determine the common traits among those individuals with a high probability to default on their credit payments. Substantiate your answer with specific reference to the output generated in the IDE notebook.
  3. Identify and briefly discuss other comparisons you would have chosen to visualise when making this prediction.

Your submission, excluding in-text citations and your list of references, may not exceed **500 words**.

Start writing here:

1.1

The credit data set likely includes categorical variables (e.g., loan approval status, gender) and numerical variables (e.g., age, income).

For numerical data, histograms are used to show the distribution of variables like income and age, which helps in understanding the central tendency and spread of these variables. Scatter plots are appropriate for examining relationships between numerical variables, such as income vs. age.

Bar charts are suitable for categorical data to compare the frequency or proportion of categories, such as the count of approved vs. declined loans.

Box plots can be used to display the distribution of numerical variables across different categories (e.g., income across different loan statuses), highlighting medians and variability.

1.2

* **Persona for Loan Approval:**
  + **Age**: 30-45 years
  + **Income**: High (above median income level)
  + **Credit Score**: High (good credit history)
  + **Employment**: Stable employment history
  + **Debt-to-Income Ratio**: Low
  + **Loan Amount**: Within a reasonable range based on income and existing debts
* **Persona for Loan Decline:**
  + **Age**: 18-25 years or above 60 years
  + **Income**: Low (below median income level)
  + **Credit Score**: Low (poor credit history)
  + **Employment**: Unstable or short employment history
  + **Debt-to-Income Ratio**: High
  + **Loan Amount**: Higher than what their income and debt levels would reasonably allow

1.3

* **Credit Utilization vs. Loan Approval**: A scatter plot showing credit utilization rates against loan approval status could reveal whether individuals with higher credit utilization are more likely to be declined.
* **Debt-to-Income Ratio vs. Loan Status**: This could help in understanding how much debt relative to income affects the probability of loan approval.
* **Employment Duration vs. Loan Status**: A bar chart comparing the employment duration of those who were approved versus those who were declined.
* **Loan Amount vs. Approval Status**: A box plot showing the distribution of loan amounts for approved and declined applications could provide insights into whether higher loan amounts are linked to higher declination rates.

#### 4. Rubric

The following rubric will be used to grade your submission for this activity submission.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Unsatisfactory** | **Limited** | **Accomplished** | **Exceptional** |
| **Adherence to the brief**  *The student creates a persona for both the customers they would typically award and decline credit.*  *The student substantiates their answer based on the output of the IDE notebook.* | Answer fails to adhere to any of the elements contained in the brief. (0) | Answer adheres to some elements contained in the brief, but some key elements are missing. (2) | Answer adheres to most, but not all, elements of the brief. Almost all information is provided and relevant. (4) | Answer adheres to all the elements of the brief. All information provided is comprehensive and relevant. (6) |
| **Evidence of understanding and accurate use of the module’s content**  *The answer demonstrates that the student has engaged with the content.*  *The answer demonstrates that the student has an informed grasp of the types of data, and how data is cleaned, imported into R, and used to create plots to inform a business decision.* | There is no evidence that the student has engaged with the content.  OR  The student fails to demonstrate a basic understanding of the content. (0) | There is little evidence that the student has engaged with the content. The understanding that is evident is inadequate. (2) | There is evidence that the student has engaged with the content and understands most of it. (4) | The student has an excellent understanding of the module’s content. (6) |
| **Coherence and clarity**  *The answer is clearly structured and written in a way that is comprehensible.*  *The student adheres to the word count.* | Answer is incoherent or lacks clarity. Answer is not logically structured or is incomprehensible. (0) | Answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. Answer does not fall within the prescribed word count (50 words over the word count). (2) | Answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. Answer falls within the word count. (4) | Answer is extremely well-structured and written with exceptional clarity and coherence. Answer falls within the word count. (6) |

**Total:** 18 marks

#### Feedback

Start writing here:

#### References

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An introduction to statistical learning with applications in R.* New York: Springer-Verlag.

Paisa Bazaar. n.d. *Credit rating in India.* Available: [https://www.paisabazaar.com/cibil/credit-rating](https://www.paisabazaar.com/cibil/credit-rating/) [2019, November 18].