

ANL252

End-of-Course Assessment - January Semester 2023

Python for Data Analytics

INSTRUCTIONS TO STUDENTS:

1. This End-of-Course Assessment paper comprises **6** pages (including the cover page).
2. You are to include the following particulars in your submission: Course Code, Title of the ECA, SUSS PI No., Your Name, and Submission Date.
3. Late submission will be subjected to the mark deduction scheme. Please refer to the Student Handbook for details.

IMPORTANT NOTE

ECA Submission Deadline: Monday, 06 March 2023 12:00 pm

ECA Submission Guidelines

Please follow the submission instructions stated below:

A - What Must Be Submitted

You are required to submit the following item for marking and grading:

- *A Report*

Please verify your submissions after you have submitted the above item.

B - Submission Deadline

- *The Report is to be submitted **by 12 noon** on the submission deadline.*
- *You are allowed multiple submissions till the cut-off date.*
- *Late submission **will be subjected to mark-deduction scheme** by the University. Please refer to Section 5.2 Para 2.4 of the Student Handbook.*

C - How the (1) Item Should Be Submitted

- *The Report: submit online to Canvas via TurnItIn (for plagiarism detection)*
- *Avoid using a public WiFi connection for submitting large files. If you are using public wireless (WiFi) connection (e.g. SG Wireless at public areas), you might encounter a break in the connection when sending large files.*

D - Additional guidelines on file formatting are given as follows:

<p>1. Report</p>	<ul style="list-style-type: none"> • Please ensure that your Microsoft Word document is generated by Microsoft Word 2007 or higher. • The report must be saved in .docx format. • The ECA is to be answered using Python and any related libraries. One may investigate and use any suitable functions/methods in the Numpy or Pandas libraries. There is no restriction in using any other built-in Python functions methods or other Python libraries, if necessary. • The complete Python codes indicated clearly corresponding to each part of the question, are to be expressed in text format and must be included, with the correct indentation(s), in the appendix of the report. Screenshots of the codes are not permitted and will not be marked. • The charts produced are to be included as images in the word document. • The context of the dataset relates to some corporate claims processing of an insurance company. The data dictionary of this dataset is depicted in Appendix 1.
-------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

E – Please be Aware of the Following:

Submission in hardcopy or any other means not given in the above guidelines will not be accepted. You do not need to submit any other forms or cover sheets (e.g. form ET3) with your ECA.

*You are reminded that electronic transmission is not immediate. The network traffic may be particularly heavy on the date of submission deadline and connections to the system cannot be guaranteed. Hence, you are advised to submit your work early. **Canvas will allow you to submit your work late but your work will be subjected to the mark-deduction scheme.** You should therefore not jeopardise your course result by submitting your ECA at the last minute.*

It is your responsibility to check and ensure that your files are successfully submitted to Canvas.

F - Plagiarism and Collusion

Plagiarism and collusion are forms of cheating and are not acceptable in any form in a student's work, including this ECA. Plagiarism and collusion are taking work done by others or work done together with others respectively and passing it off as your own. You can avoid plagiarism by giving appropriate references when you use other people's ideas, words or pictures (including diagrams). Refer to the APA Manual if you need reminding about quoting and referencing. You can avoid collusion by ensuring that your submission is based on your own individual effort.

The electronic submission of your ECA will be screened by plagiarism detection software. For more information about plagiarism and collusion, you should refer to the Student Handbook (Section 5.2.1.3). You are reminded that SUSS takes a tough stance against plagiarism or collusion. Serious cases will normally result in the student being referred to SUSS's Student Disciplinary Group. For other cases, significant mark penalties or expulsion from the course will be imposed.

(Full marks: 100)

Section A (100 marks)

Answer all questions in this section.

Question 1

Besides blanks, 'Unkn' and '???' are expressions in the dataset that denote missing values. With this consideration, read in the dataset as a Pandas dataframe, and state the variables that contain missing values.

(5 marks)

Question 2

As part of data preparation, treat the missing data, and explain your rationale of the treatments.

(15 marks)

Question 3

Explain and implement *three (3)* other data preparation tasks required for further analysis of the data. Any appropriate Python related libraries, functions, methods (e.g. `pandas.to_datetime`) can be used.

(15 marks)

Question 4

Analyse the data and describe *three (3)* insights into the corporate claims processing of the insurance company, with at least *one (1)* supporting visualization created to illustrate each insight.

(30 marks)

Question 5

Perform linear regression modelling to predict the delay in days (between the Planned and Actual date) in processing the claims, explaining the approach taken, including any further data pre-processing needed for modelling.

(25 marks)

Question 6

Discuss the results obtained from the modelling and state the linear regression equation.

(10 marks)

Appendix:

APPENDIX 1 – DATA DICTIONARY

Variable	Description
Claim_ID	Unique identifier of claim
Policy_No	Unique identifier of corporate policy tied to an organization
Name	Name of claimant
Planned	Planned date of claim settlement
Actual	Actual date of claim settlement
Created	Claim settlement record creation date
Amount	Payout amount
Paid	Status of payment (Yes or No)
Category	Internal categorization code
Terms	Internal terms and conditions code
Region	Internal region classification code
Type	Internal type classification code

Note: The meaning of each value for the internal codes of the organization is unknown.

----- END OF ECA PAPER -----