SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

# ANL252
# Python for Data Analytics

---

## Group-based Assignment

## July 2022 Presentation

---

## GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for ANL252 Python for Data Analytics.

The cut-off date for this assignment is **21 August 2022, 2355hrs.**

This is a group-based assignment. You should form a group of **4 members** from your seminar group. Each group is required to upload a single report via your respective seminar group site in Canvas. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment. If you feel that the work distribution is inequitable to either yourself or your group mates, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective.

Note to Students:

You are to include the following particulars in your submission: Course Code, Title of the GBA, SUSS PI No., Your Name, and Submission Date.

---

### Question 1

The dataset comprises records of commuter journeys. The data dictionary of this dataset is depicted in Appendix 1.

The Python codes are to be expressed in text format and must be included, with the correct indentation(s), in all the answers in the report. **Screenshots of the codes are not permitted and will not be marked**.

The corresponding Jupyter notebook must also be submitted.

(a) In the dataset, '-', '--', and '?' are considered as missing values, and the variable columns of the dataset are noted as in the data dictionary in order. As part of data preparation, read the dataset in as a Pandas dataframe, with the above considerations.
(3 marks)

(b) Identify the variable columns which have missing values. As part of data preparation, implement ways to treat them, and explain your rationale. State any interesting observation(s).

(18 marks)

(c) As part of data preparation, identify ***three (3)*** other data quality issues in the data. Similarly, suggest and implement ways to treat them, and explain your rationale.

(24 marks)

(d) Develop a user-defined function that will print the hour, expressed in the 12-hour clock format (e.g., 12am, 1pm), whereby the highest number of commuters start their journey.

(10 marks)

(e) Write a Python code to create appropriate visualisations of the commuter data. Analyse the results and then discuss ***three (3)*** interesting insights.

(45 marks)

## APPENDIX 1 – DATA DICTIONARY

| Variable | Description |
|---|---|
| origin | Start location identifier |
| destination | End location identifier |
| start | Start time |
| end | End time |
| id | Device/Vehicle identifier |
| type | Customer profile type |
| subscriber | Subscribing customer (Yes or No) |
| yob | Customer year of birth |
| age | Customer age |
| gender | Customer gender |

**---- END OF ASSIGNMENT ----**