# MELBOURNE BUSINESS SCHOOL (The University of Melbourne)
# MBUSA90500 Syndicate Programming Assignment 2022 V1.0

## Due date 23:59 5th July.
## Syndicate presentation 7th July.

## 1 Objectives

This project is designed to give your syndicate experience in software engineering, Python programming and processing data from different sources. In particular, you should experience the following.

- Python programming - every member of the syndicate must contribute code to the final solution.
- Software design - the process of breaking a task into modules and assigning each module to a programmer in such a way that the pieces can be easily assembled.
- The frustration of unclean data sets.
- Software testing - each module should be tested prior to integration, and then the whole tested once integrated.

## 2 The Problem

The City of Melbourne has been hard hit over the last 2 years due to the COVID-19 pandemic and the amount of pedestrian traffic has changed considerably. Your company *Modellers-R-Us* specialises in modelling pedestrian activity in parts of the city of Melbourne. You are developing Python code that will help analyse pedestrian activity across different parts of the City of Melbourne. In particular, your software will analyse pedestrian counts and relation to other factors, such as time, maximum temperature, rainfall and solar exposure.

## 3 Data

To tackle this project, you will use the following open datasets.

**Pedestrian counts from sensors in the City of Melbourne 1/1/2021 to 31/5/2022**

- count2021-2022.csv
- The source data website is here.

**Melbourne weather data: maximum temperature (in Celsius), rainfall (in mm) and solar exposure (in MJ/mm$^2$) 1/1/2021 to 31/5/2022**

- temperature-all-years.zip, rainfall-all-years.zip, solar-all-years.zip
- The source data website is here
- Measurements come from the Melbourne Olympic Park Weather Station.
  - Daily rainfall
  - Daily maximum temperature
  - Daily solar exposure
- You will need to extract data for the period 1/1/2021-31/5/2022 from these files.

# 4 Analysis

Using these open datasets, you will need to write Python code to perform analysis for each of the following 15 questions. You may use any Python libraries that you wish. For questions which require plots, you may wish to investigate libraries such as matplotlib or seaborn. Note that we are not expecting you to produce fancy graphics, just basic plots which display the information requested. You should ensure that all plots are clearly labelled.
In the descriptions below, assume unless otherwise specified that

- By 2021, we mean the period 1/1/2021-31/12/2021
- By 2022, we mean the period 1/1/2022-31/05/2022

## 4.1 Questions

1. Divide each weekday into three periods consisting of a morning rush hour (8:00–9:00), lunch hour (13:00–14:00), and evening rush hour (17:00–18:00). Each sensor can be associated with a count for each of these periods. Summing over all sensors we have a period's overall pedestrian count on a particular week day. Compute the following statistics for each of the three periods in 2022.
    - Mean, median, min, max of the period's overall pedestrian count on weekdays
2. Let the daily overall pedestrian count be the total number of pedestrians counted by sensors on a particular day. For each of 2021, 2022, generate a scatter plot of maximum temperature (x axis) and daily overall pedestrian count (y axis)
3. For each of 2021, 2022, generate a scatter plot of rainfall (x axis) and daily overall pedestrian count (y axis)
4. For each of 2021, 2022, generate a scatter plot of solar exposure (x axis) and daily overall pedestrian count (y axis)
5. For each of 2021, 2022, generate a histogram showing how busy (mean daily overall pedestrian count y axis) each day of the week (x axis) is.
6. For 2022, generate a histogram showing the mean daily overall pedestrian count (y axis) for sensors 1-20 (x axis).
7. Same as Q5, but in the calculation only include days where it rained.
8. Same as Q5, but in the calculation only include days where the maximum temperature was less than 20 and it rained.
9. Consider the time series of the daily overall pedestrian count for each sensor in May 2021 and in May 2022. Comparing these two months, which sensor's time series changed the most and by how much? [you should use Euclidean distance to compare two time series]
10. Suppose we wish to predict the pedestrian count recorded by sensor 3 Melbourne Central during 12-1pm for any particular day. One may model this as a regression problem and fit a linear regression model to estimate this count. Use data from January-April 2022 to fit your model and data from May 2022 to evaluate its performance. You may decide yourself which variables to use as independent variables, but you should include at least 5 independent variables and some ideas are:
    - Pedestrian counts from this sensor or other nearby sensors in previous hours.
    - Temperature from the previous day
    - Solar exposure from the previous day
    - Rainfall from the previous day
    - ...

    Whatever variables you choose, you should ensure that you don't rely on using any information that would not be available before the 12-1pm period on a given day. Estimate the performance of your model using an appropriate metric. For what day of the week is it most accurate, for what day of the week is it least accurate?
11. For sensor 3 Melbourne Central, return the top 3 most unusual days of pedestrian activity recorded by this sensor during 2022. For each of these three days, generate a scatter plot of the pedestrian counts from this sensor across the day. [you will need to make a choice here about how to assess "unusualness". There is no best answer here, but you will need to be able to justify your choice].

12. During May 2022, on which day of this month is the pedestrian traffic recorded by (sensor 9 Southern Cross Station) and (sensor 3 Melbourne Central) most similar/different? [similar to Q9, use Euclidean distance to compare two time series. In this case, the time series of counts across a day]

13. One may compare the time series of counts from two sensors, by calculating the Pearson correlation between them. Consider the time period 0900-1700, on which weekday during May 2022, do (sensor 9 Southern Cross Station) and (sensor 3 Melbourne Central) have i) the highest (absolute) correlation?; ii) the lowest (absolute) correlation?

14. An anomalous local event is a short lived event which occurs in a single location of the city (or a small set of spatially close neighbor locations) and is reflected by an unexpected change in pedestrian count. Identify one anomalous local event that happened during 2021 and one that happened during 2022. [you will need to make a choice here about how to detect anomalous local events. There is no single right answer here, but you will need to be able to justify your strategy].

15. Combine some other open dataset with the data you have been provided with. Demonstrate the value added by your additional data with an appropriate plot or visualisation.

You may need to make some assumptions about the data to answer these questions. You should state these assumptions in the report.

# 5 Report

Prepare a short report containing the answers to each of the XXX questions above, based on your Python code. For each of the 15 questions,

1. Show the results in a table (4 statistics for each time period. $4 * 3 = 12$ statistics in total)
2. Show the 2 scatter plots
3. Show the 2 scatter plots
4. Show the 2 scatter plots
5. Show the 2 histograms
6. Show the histogram
7. Show the 2 histograms
8. Show the two histograms
9. State which sensor and how much the change was. Provide an explanation for why it is reasonable that this sensor changed more than any other sensor.
10. Explain what variables were used by your regression model. State how much data was used to train (fit) your model and how much was used to test the performance of your model. Provide performance measures for your regression model. State for what day of the week it is most accurate and for what day of the week it is least accurate, providing performance numbers.
11. State which 3 days and show the 3 requested scatter plots. Explain the approach you used to measure "unusualness and why it is reasonable. Comment on your scatter plots and argue why they are reasonable.
12. State on which day they are most similar, how much the difference is and comment on your result. State on which day they are most different, how much the difference is and comment on your result.
13. State the weekday in May 2022 with the highest absolute correlation and the weekday in May 2022 with the lowest absolute correlation. Comment on your results.
14. List two anomalous events, one in 2021 and one in 2022 (for each event, the sensor(s) and time period). Provide accompanying visualisation to support why they should be considered anomalous. Speculate on what caused these events.
15. Describe the extra data and its characteristics. Showcase the value it adds via an appropriate plot or visualisation.

**Assumptions:** You should also include a section of the report "Assumptions", which details any assumptions you needed to make and why.
**Data cleaning:** You should include a section of the report "Data cleaning", which details any steps you took to pre-process the data and why (e.g. remove errors perform imputation).

# 6 Presentation

You will be required to give a short (7min presentation). You will need to cover

- Your approach and accompanying results for Q10, Q11, Q14 and Q15
- Considering all 15 questions, what have been the challenges and what (if anything) would you have done differently in retrospect?

# 7 Assessment

## 7.1 Within-syndicate responsibilities

We expect that you will structure the Python code so that it is split into modules/functions, and different members of the syndicate will write different parts of the code. Naturally you can collaborate, but it is in the best interests of the less experienced coders that they be given a coding task to complete. The temptation for the stronger coders to do the lot is high, but there will be coding questions about this assignment on the exam that each individual must answer. We suggest the stronger coders concentrate on the design, and splitting the tasks up amongst the syndicate, aiming to allot the tasks based on a syndicate member's abilities.

## 7.2 Submission and Presentation

Your submission should have the following format.

- The submission should be a single zip file (**not** tar, rar, xz, ...) for the syndicate
- The zip file should contain
  - A pdf document (not Word) providing the information requested for the Report Section.
  - All of the Python code that implements the required functions.
  - A README.txt file that provides instructions on how to run your code. and a description of how your Python is structured. What are the main modules/files/functions and what do they do? There is no need to describe trivial functions, just the main idea of your software. Perhaps a diagram? Also include instructions on how we should run it.
  - A CONTRIBUTION.txt file listing which syndicate members wrote which python code.

The submission must be via Canvas.

Your submission should be before 23:59 5th July.

You will be required to give a short 7 minute presentation in class on the afternoon of 7th July.

## 7.3 Timeline

| **Item** | **Due date** | **Submission method** |
|---|---|---|
| Submission of data and code | before 23:59 5 July | via Canvas |
| Short 7 min presentation in class | afternoon of 7 July | in workshop |

## 7.4 Marking Rubric for Syndicate

| Component | Property | Mark |
|---|---|---|
| Python coding | Parsing of csv files | /3 |
| (12 marks) | Code is Clean, Commented Readable/Maintainable | /6 |
| | Some evidence of code testing[1] | /3 |
| Analysis answers | All syndicate's answers correct | 6 |
| (6 marks) | Most syndicate's answers correct | 4 |
| | Not many syndicate's data correct | 2 |
| | Doesn't work | 0 |
| Report | Answers to each question clearly presented and plots clearly labelled | 4 |
| (11 marks) | Appropriate and clear explanation/justification provided for methods used and their rationale (Q10,Q11,Q14,Q15). | 6 |
| | Assumptions and data cleaning sections are clear and appropriate | 1 |
| Presentation | Clear descriptions of requested items | 1 |
| (5 marks) | Challenges described | 1 |
| | Effective use of visual resources | 1 |
| | Handling of questions | 1 |
| | Adhered to time | 1 |
| | Total | /34 |

* For example, functions called `test_xxx()` or even a separate testing modules `test.py`.