



NGEE ANN
P O L Y T E C H N I C

School of InfoComm Technology

Data Wrangling

Diploma in Data Science

INDIVIDUAL ASSIGNMENT II

(40% of Data Wrangling Module)

03 Jul 2023 – 19 Aug 2023

Deadline for Submission:

**Jupyter Notebook File and Powerpoint Slides:
19 Aug 2023 (Sat), 2359hrs**

Student Name	:
Student Number	:

Penalty for late submission:

10% of the marks will be deducted every day after the deadline.

NO submission will be accepted after **26 Aug 2023, 23:59**.

DATA WRANGLING ASSIGNMENT 2

1. OBJECTIVES

In this assignment we will extract the data from a real-life database, wrangle and prepare the data to solve a prediction problem. (regression, classification)

- To extract data from a database, explore the data and formulate a prediction problem
- To create a tabular data table from multiple tables based on the formulated problem
- To wrangle and prepare the data ready for modeling, use the prepared data to build and evaluate a simple machine learning model
- To document the process, analysis, comparison and findings

2. DATASET: F1 DATABASE FROM ERGAST

Ergast.com is a webservice that provides a database of Formula 1 races, starting from the 1950 season until today. The dataset includes information such as the time taken in each lap, the time taken for pit stops, the performance in the qualifying rounds etc. of all Formula 1 races.

You can download the **datasets.zip** file from Politemall, where you can find:

- **f1_db_csv** folder: a total of 13 .csv files / tables.
- **f1db_data_dictionary.txt** file: detailed description and information for all the 13 tables.

You should load data from the CSV files for use in ASG2.

If you would like to understand more context about F1, please refer to this Wikipedia website https://en.wikipedia.org/wiki/Formula_One_racing.

3. SUGGESTED TASKS

You are suggested to complete this assignment following the below steps.

ALL THE STEPS ARE REQUIRED TO BE DONE THROUGH PYTHON IN JUPYTER NOTEBOOK.

Step 1: Problem Formulation

Load the data from CSV files. Explore the data, understand the data and formulate a prediction problem. It can be a regression problem or classification problem and you need to utilize the information from at least **THREE** different tables to solve this problem.

Step 2: Data Wrangling on multiple tables

Based on the formulated prediction problem, create a Tabular Data table by extracting data from multiple tables. You may need to utilize the below techniques in this step:

- Subsetting, Grouping and Filtering the tables
- Concatenation, Merging and Joining the tables
- Create features with Transactional Data or Time Series Data
- Applying Mathematical Calculations to features
- Extract features from unstructured data (e.g. Text data, Data and Time and etc.)

Step 3: Data Cleansing and Transformation

Cleanse and transform the tabular data before feed it into the Machine Learning models. You may need to utilize the below techniques in this step:

- Missing value imputation
- Outliers removal/capping
- Categorical Data Encoding
- Numerical Data Transformation
- Variable Binning or Discretization
- Feature Scaling
- Applying Mathematical Calculations to features

Step 4: Machine Learning Modelling

State number of rows and columns in your final dataset before building machine learning models. This will help show that your predictions are not trivial nor unrealistic, eg. 100% accuracy when predicting on total of 5 rows of data only, or perhaps having extremely little number of X columns (1 – 2), despite the wealth of data on hand.

Build both a naïve baseline and a simple machine learning model and evaluate the model performance. Are you happy with the model performance? If not, please review the previous steps and see whether you can further wrangle the data to improve the model performance.

4. SUGGESTED REPORT FORMAT & CONTENT GUIDELINES (TO BE INCORPORATED INTO JUPYTER NOTEBOOK)

Write an accompanying **INDIVIDUAL** report with the following sections within your Jupyter Notebook file, using Markdown cells (see Table below). Please have the report at the bottom of your Jupyter Notebook, you are free to paragraph and/or section as necessary.

You can refer to this quick guide on using and writing reports and commentary with Markdown in Jupyter Notebook:

<https://www.datacamp.com/community/tutorials/markdown-in-jupyter-notebook>

Sample content is provided for each section. You are free to include other relevant information you deem necessary in the sections. **You are strongly encouraged to try different methods at each section and provide detailed comparison and discussion in the report.**

	Suggested Report Sections & Content Guidelines	Recommended Word Count
1.	Table of Contents	100 – 500 words
2.	Introduction with Value Based Problem Statement	250 – 500 words
3.	Problem Formulation <ul style="list-style-type: none"> • Load and Explore the Data • Understand the Data • Formulate a Prediction Problem 	250 – 500 words
4.	Data Wrangling on multiple tables <ul style="list-style-type: none"> • Extract and Create features from different tables • Concatenate, Merge or Join the tables 	250 – 500 words
5.	Data Cleansing and Transformation <ul style="list-style-type: none"> • Missing Value and Outliers • Categorical Data • Numerical Data • Others 	250 – 500 words
6.	Machine Learning Model <ul style="list-style-type: none"> • Show Count of Rows and Columns • Build and Evaluate the model against a Naïve Baseline Model 	250 – 500 words
7.	Summary and Further Improvements <ul style="list-style-type: none"> • Summarize your findings • Explain the possible further improvements 	100 – 500 words

5. DELIVERABLES

Presentation and demonstration

- Each student is required to do an **online live presentation** and share your findings. **The presentation should not exceed 10 minutes.** The presentations which exceed the allotted time will be penalized.
- **Students are to book one-to-one presentation timeslots** *scheduled by your lecturer* during **Week 17 and 18's regular lesson date.**
- Students are to submit the presentation slides that is used for the Presentation in Politemall. Deadline for slides submission is **Sat, 19 Aug 2023, 2359 hours.**

Assignment files

- Submit the Jupyter Notebook file (DW_ASG2_InsertStudentName.ipynb) and Powerpoint Slides (DW_ASG2_InsertStudentName.pptx) in a zipped format in Politemall. Deadline for submission is **Sat, 19 Aug 2023, 2359 hours.**
- Run-time errors will result in significant marks penalties, please fully rerun your notebook successfully before submission.

Note: DO NOT PLAGIARIZE (<https://www1.np.edu.sg/clte/antiplagiarism/policy.htm> for more information)

6. GRADING CRITERIA

		<u>Quality of Presentation Slides (10 marks)</u> Assessed based on: <ul style="list-style-type: none"> design of slides, and effective use of visualizations proper use of appropriate vocabulary, and conciseness slides to be free from spelling errors, leftover template artefacts, etc. 	<u>Presentation Skills (10 marks)</u> Assessed based on: <ul style="list-style-type: none"> whether the presentations are clear, concise and well-organized whether presenters show clear understanding of work done meeting typical video presentation norms (video on, adequate sound level, etc.)
<u>Quality of Work, Report (30 marks)</u> Assessed based on: <ul style="list-style-type: none"> work showing depth and quality of the business problem based on the given dataset work showing good rationale and considerations of datasets and wrangling techniques chosen 	<u>Completeness of Report Based on Content Guidelines (30 marks)</u> Assessed based on: <ul style="list-style-type: none"> strong narrative of steps taken during data wrangling breadth of steps as per section 4, that multiple approaches are tested for each DW step detection of any errors found, with executed correction done well 	<u>Analysis and Discussion (10 marks)</u> Assessed based on: <ul style="list-style-type: none"> showcasing good conclusions from work done discussion on steps taken, with explanation of various degrees of success clear, logical explanations, throughout the report 	<u>Report Writing (10 marks)</u> Assessed based on: <ul style="list-style-type: none"> formatting of report, and effective use of visualizations proper use of appropriate vocabulary, and conciseness report to be at bottom of Jupyter Notebook, though additional comments throughout the notebook is fine.