NGEE ANN POLYTECHNIC

School of InfoComm Technology

Data Wrangling

Diploma in Data Science

INDIVIDUAL ASSIGNMENT I

(30% of Data Wrangling Module)

01 May 2023 – 03 Jun 2023

Deadline for Submission:

Jupyter Notebook File and Powerpoint Slides:
03 Jun 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 10 Jun 2023, 23:59.

DATA WRANGLING ASSIGNMENT 1

1. OBJECTIVES

In this assignment we will wrangle the data from a real-life dataset to understand different data wrangling techniques.

- To conduct data exploration, preparation and transformation through different methods
- To prepare the data ready for modeling, build and evaluate a simple linear regression model.
- To document the analysis, comparison and findings

2. DATASET: SUPERMARKET SALES FORECAST (REGRESSION PROBLEM)

The data ('**supermarket.csv**') have been collected at various supermarket outlets and stores in different cities. The aim is to predict the sales of each product at a particular outlet. Using this, supermarket management team will try to understand the properties of products and outlets which play a key role in increasing sales.

Detailed information (i.e. column description) is provided below.

- **Item_Weight:** Weight of product
- **Item_Fat_Content:** Whether the product is low fat or not
- **Item_Visibility:** The % of total display area of all products in a store allocated to the particular product
- **Item_Type:** The category to which the product belongs
- **Item_MRP:** Maximum Retail Price (list price) of the product
- **Outlet_Identifier:** Unique store ID
- **Outlet_Establishment_Year:** The year in which store was established
- **Outlet_Size:** The size of the store in terms of ground area covered
- **Outlet_Location_Type:** The type of city in which the store is located
- **Outlet_Type:** Whether the outlet is just a grocery store or some sort of supermarket
- **Item_Outlet_Sales:** Sales dollar amount of the product in the particular store. This is the TARGET variable.

3. SUGGESTED TASKS

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

Step 0: **Exploratory Data Analysis (EDA)**

Download the dataset from Politemall, conduct exploratory data analysis using TIBCO Spotfire. Investigate the relationships between different features/variables. Which features are likely helpful for making predications?

ALL THE BELOW STEPS WILL BE DONE THROUGH PYTHON.

Step 1: **Load Data into Jupyter Notebook via Relative Filepath(s)**
Load the data into a DataFrame variable and provide an overview of the DataFrame variable using the relevant functions(e.g. head(), info(), describe() and etc.)

Step 2: **Data Preprocessing**
Are there any outliers? How did you identify them and how to deal with them? Are you happy with the distribution of the numerical variables? Do you need to transform the numerical variables using proper transformation methods (e.g. log transformation, Box-Cox and etc.)?

Step 3: **Train and Test Split**
Split the data into train data (70%) and test data (30%)

Step 4: **Missing Value Imputation**
Are there any missing values? How did you handle them and why?

Step 5: **Categorical Data Encoding**
Do you need to encode the Categorical Data? What methods do you use and why?

Step 6: **Variable Discretization /Binning**
Do you need to discretize /bin the Numerical Data? What methods do you use and why?

Step 7: **Feature Engineer**
Do you need to scale the data? What method do you use and why? Do you create any new features/variables and why? Do you drop any features/variables and why?

Step 8: **Linear Regression Modelling**
Use the sample code provided in the jupyter notebook file for assignment submission to execute this part:

> Build a linear regression model and evaluate the model performance. Are you happy with the model performance? If not, please review the previous steps 2-7 and see whether you can further wrangle the data to improve the model performance.

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

Write an accompanying **INDIVIDUAL** report with the following sections within your Jupyter Notebook file, using Markdown cells (see Table below).

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.	Introduction: Problem Understanding	100 – 500 words
2.	Explore the Data • the relationship between different variables / features	250 – 500 words
3.	Cleanse the Data • Missing Data • Outliers	250 – 500 words
4.	Data Transformation • Categorical Data (e.g. One hot encoding, Ordinal label encoding and etc.) • Numerical Data (e.g. log transformation, binning)	250 – 500 words
5.	Feature Engineer • Feature Scaling • Create new features /Drop features	250 – 500 words
6.	Linear Regression Model • Build and Evaluate the model	250 – 500 words
7.	Summary and Further Improvements • Summarize your findings • Explain the possible further improvements	100 – 500 words

5. DELIVERABLES

Presentation and demonstration

- Record presentation using any recording tool (such as MS Teams etc)
- Upload the recorded video as “unlisted” YouTube to be viewable by the tutor as outlined in this URL: https://youtu.be/WkgOvUr5Alc. For more information on the video recording and how to upload as unlisted YouTube, please refer to the reference material “Video Recording for Assignment”.
- You are required to do an online presentation and share your findings. The presentation should not exceed 10 minutes. The presentations which exceed the allotted time will be penalized.
- Prepare and format slides for your presentation.
- Copy and paste the YouTube video link in final cell of Jupyter Notebook. **Make sure your lecturer is given rights to view.** Students are to submit the presentation slides that are used for the Presentation in Politemall. Deadline for slides submission is **Sat, 03 Jun 2023, 2359 hours.**

Assignment files

- Submit the Jupyter Notebook file (DW_ASG1_InsertStudentName.ipynb) and Powerpoint Slides (DW_ASG1_InsertStudentName.pptx) in a zipped format in Politemall. Deadline for submission is **Sat, 03 Jun 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

		Quality of Presentation Slides (10 marks) Assessed based on: • 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.	**Presentation Skills (10 marks)** Assessed based on: • 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.)
Quality of Work, Report (30 marks) Assessed based on: • 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	**Completeness of Report Based on Content Guidelines (30 marks)** Assessed based on: • 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	**Analysis and Discussion (10 marks)** Assessed based on: • showcasing good conclusions from work done • discussion on steps taken, with explanation of various degrees of success • clear, logical explanations, throughout the report	**Report Writing (10 marks)** Assessed based on: • 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.