ngee ann school of infocomm technology

Data Wrangling

Diploma in Data Science

INDIVIDUAL ASSIGNMENT II

(50% of Data Wrangling Module)

1 Jul 2024 – 11 Aug 2024

Deadline for Submission:

Slides: 28 Jul 2024 (Sun), 2359hrs

Report: 11 Aug 2024 (Sun), 2359hrs

Student Name	:	Tan Jun Yu Zavier
Student Number	:	S10255651D

Penalty for late submission:

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

NO submission will be accepted after **18 Aug 2024, 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.

- 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: THE HISTORY OF BASEBALL

This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2014. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875.

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

- CSV folder: a total of 26 .csv files / tables.
- **data_dictionary.doc** file: detailed description and information for all the 26 tables.

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

3. SUGGESTED TASKS

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

ALL THE STEPS ARE REQUIED 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 a 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 etc.)

Step 3: Data Cleansing and Transformation

Cleanse and transform the tabular data before feeding 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

Build both a naïve baseline and a simple machine learning model and evaluate the model performance. Are you happy with the model's 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 (INCORPORATE 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	**Word Count**
1.	Table of Contents	NA
2.	Introduction with Value Based Problem Statement	Min: 100 words Max: 500 words
3.	Problem Formulation • Load and Explore the Data • Understand the Data • Formulate a Prediction Problem	Min: 500 words Max: 1500 words
4.	Data Wrangling on multiple tables • Extract and Create features from different tables • Concatenate, Merge or Join the tables	Min: 1000 words Max: 2000 words
5.	Data Cleansing and Transformation • Missing Value and Outliers • Categorical Data • Numerical Data • Others	Min: 1000 words Max: 2000 words
6.	Machine Learning Model • Build and evaluate the model against a Naïve Baseline Model	Min: 500 words Max: 1000 words
7.	Summary and Further Improvements • Summarize your findings • Explain the possible further improvements	Min: 100 words Max: 500 words
8.	Reflection **With reference to the module learning objectives, reflect on the knowledge, skills, and abilities learnt and any areas which you could have improved upon your own learning process.** Our module learning objectives are as follows: • Understand the Data Wrangling process and principles. • Extract and transform features from structured datasets, time series, transactions data and text. • Design data pipelines to do categorical variable encoding, numerical variable transformation, missing data imputation and variable discretization. • Utilize the cutting-edge Python packages to extract, transform and generate features to develop highly optimized machine learning models. • Utilize the enterprise software to do inline data wrangling. • Apply the data wrangling techniques to solve real world problems.	Min: 500 words Max: 1000 words

5. DELIVERABLES

Presentation and demonstration

- Each student is required to do a **face-to-face live presentation** and share their findings. **The presentation should not exceed 8 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 16 and 17's regular lesson date**.
- Students are to submit the **presentation slides** that is used for the Presentation in Politemall. Deadline for slides submission is **Sun, 28 Jul 2024, 2359 hours.**

Assignment report

- Submit the **Jupyter Notebook file** (DW_ASG2_InsertStudentName.ipynb) containing your codes and report through PoliteMall. Deadline for submission is **Sun, 11 Aug 2024, 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 chose	**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 • Reflections in Jupyter Notebook are clear and well-written.