Data Wrangling

Diploma in Data Science

INDIVIDUAL ASSIGNMENT I

(30% of Data Wrangling Module)

06 May 2024 – 01 Jun 2024

Deadline for Submission:
Jupyter Notebook File,
Video and Powerpoint Slides:
01 Jun 2024 (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 08 Jun 2024, 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: SONG'S POPULARITY (REGRESSION PROBLEM)

As a data analyst at XYZ Music Records Company, you have been tasked with investigating the influence of various song characteristics on a song's popularity using the dataset 'song_popularity.csv'. The dataset includes an incomplete data dictionary as given in Table 1. Your analysis aims to **predict a song's popularity** based on its various attributes. This analysis will offer valuable insights to the management team, aiding them in optimizing their allocation of marketing resources.

The available data dictionary is provided below.

Song_Name	Name of the song
Song_Popularity	Song Popularity
Song_Duration_ms	Song Duration (ms)
Acousticness	Measures the extent to which the music is made with acoustic instruments
Danceability	Quantifies how suitable a track is for dancing
Energy	Energy of song
Instrumentalness	Measures the extent to which the track is composed of instrumental sounds
Speechiness	Quantifies the degree to which spoken word elements are present in a song.
Tempo	Tempo of song

Table 1 Data Dictionary for song_popularity.csv

3. SUGGESTED TASKS

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

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: **Exploratory Data Analysis (EDA)**
Download the dataset from PoliteMall, conduct a comprehensive exploratory data analysis on the dataset.

Step 3: **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 4: **Train and Test Split**
Split the data into train data (70%) and test data (30%)

Step 5: **Missing Value Imputation**
Are there any missing values? How did you handle them and why?

Step 6: **Categorical Data Encoding**
Do you need to encode the Categorical Data? What methods do you use and why?

Step 7: **Variable Discretization /Binning**
Do you need to discretize /bin the Numerical Data? What methods do you use and why?

Step 8: **Feature Engineering**
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 9: **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). 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.	Introduction with Value Based Problem Statement; Problem Understanding	Min: 100 words Max: 500 words
2.	Exploratory Data Analysis • Insights and Conclusion	Min: 500 words Max: 1000 words
3.	Cleanse the Data • Missing Data • Outliers	Min: 500 words Max: 1000 words
4.	Data Transformation • Categorical Data (e.g. One hot encoding, Ordinal label encoding and etc.) • Numerical Data (e.g. log transformation, binning)	Min: 500 words Max: 1000 words
5.	Feature Engineer • Feature Scaling • Create new features /Drop features	Min: 500 words Max: 1000 words
6.	Linear Regression Model • Build and Evaluate the model	Min: 250 words Max: 1000 words
7.	Summary and Further Improvements • Summarize your findings • Provide recommendation for further improvements	Min: 100 words Max: 500 words

5. DELIVERABLES

Each student is required to submit the following:

- Jupyter Notebook file (with **Codes** and **Report** bottom of your Jupyter Notebook)
  - Run-time errors will result in significant marks penalties, please fully rerun your notebook successfully before submission.
- Presentation Slides
- Presentation Video (through Bongo)
  - Refer to Section 5.1 on steps for Presentation Video submission through Bongo

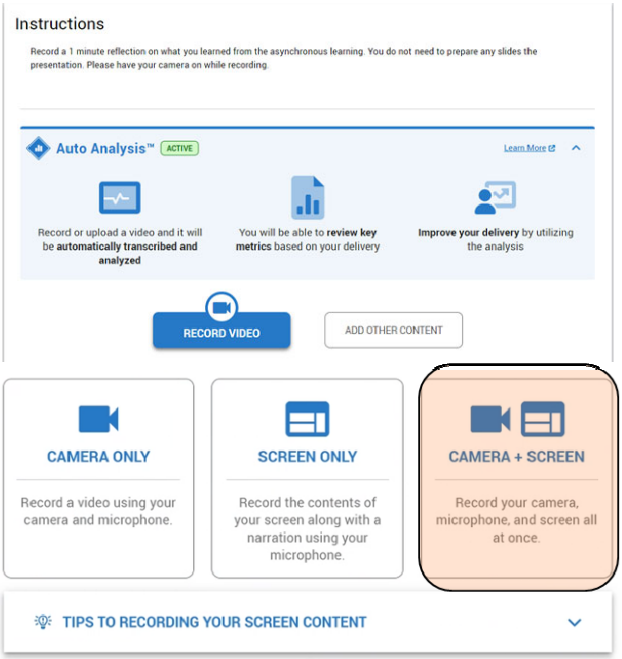
Submit the **Jupyter Notebook** file (DW_ASG1_InsertStudentName.ipynb) and **Powerpoint Slides** (DW_ASG1_InsertStudentName.pptx) in a **zipped format, via PoliteMall**. Deadline for submission is **Sat, 01 Jun 2024, 2359 hours.**

Submit **all the deliverables** no later than **Sat, 01 Jun 2024, 2359 hours** in POLITEMall. Late submissions of assignment-based coursework component without leave of absence (LOA) for the module will be subjected to the late penalty.

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

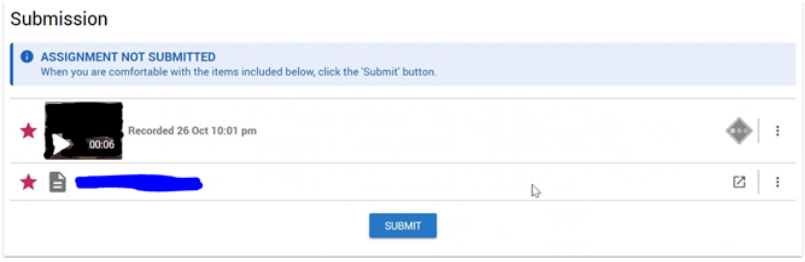
5.1. Presentation Video Submission Via Bongo

- Each student will need to record a Presentation Video of a maximum duration of **10 minutes** though the video assignment app, powered by Bongo
- Each student is to practice the presentation in advance to ensure completion **within 10 minutes**. The recording must include both webcam (clearly showing the student's face for authentication) and slides or codes (whichever is applicable).
- Select the **RECORD VIDEO** option and choose **CAMERA + SCREEN** as shown in the figure below. The figure may differ with the constantly update of the Bongo software, hence students may see a different layout but general steps should still apply.



- After recording the video, click save (as shown below) and it will be ready for students to append it for submission.

- Select the video by clicking on the Star and click **SUBMIT**.



All sections must be completed.

6. GRADING CRITERIA

	Grading Criteria	Component Weightage
Presentation	a) Quality of work b) Flow of presentation based on content guidelines (see section 4) c) Quality of presentation slides d) Presentation and articulation skills	**50%**
Practical Work & Report (Report to be incorporated to the Jupyter Notebook .ipynb file)	a) Quality of practical work b) Completeness based on suggested sections and content guidelines (see section 4) c) Quality of analysis and discussions, use of proper visual aids, d) Clarity of report, appropriate language, and grammar	**50%**