NGEE ANN POLYTECHNIC

School of InfoComm Technology

Distributed Data Pipelines

Diploma in Data Science (DS)

October 2022 Semester

INDIVIDUAL ASSIGNMENT 1

(30% of Distributed Data Pipelines Module)

Deadline for Submission:

16th Dec 2022 (Friday), 2359 Hours

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 23th Dec 2022, 23:59.

DISTRIBUTED DATA PIPELINES ASSIGNMENT 1

OBJECTIVES

The assignment aims to assess the student's understanding of the data pipelining process which include Hadoop as well as Apache Spark and PySpark.

Students will be also be tasked to execute PySpark on data sources to perform data pipelining, clean-up, and transformations, as part of building a predictive model. Student would be assessed on the following:

- Demonstration of knowledge of different stages of the data pipelining process.
- Utilization of basic components of Apache Spark via PySpark.

SECTION A

In 500 - 750 words, compare Hadoop and Apache Spark. List their similarities and differences with reference to the different stages of the data pipelining process. You are encouraged to include visual aids such as diagrams, and tables, to better illustrate your answer.

This should be done in a Word or PDF format, please refer to the later section on Deliverables, under "I".

SECTION B

1. DATASET

HDB flats spell home for 80% of Singapore's resident population, of which about 90% own their home. There are various sizes, and configurations purpose-built across the years, all over the whole island of Singapore.

Many have gotten a flat from the resale market instead of directly from HDB over the years, and these transactions have been compiled regularly by the government.

For this assignment, a modified version of some HDB transactions has been provided for a **Resale Price Prediction Model** to be built, based on the other available typical characteristics of a transaction, **via PySpark.**

The file name is **"sg_flat_prices_mod.csv"**, a single table consisting of 64248 rows, and 12 columns.

You should load data from the CSV file for use in ASG1, to **build a simple machine learning model to predict the resale prices of any given HDB resale transaction.**

2. SUGGESTED TASKS

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

ALL THE STEPS FOR SECTION B ARE REQUIRED TO BE DONE THROUGH **PYTHON IN JUPYTER NOTEBOOK VIA *PYSPARK, NO PANDAS ALLOWED*.** (please refer to the later section on Deliverables, under "II".)

Step 1: **Problem Statement Formulation**

Load the data from the CSV file. Explore the data, understand the data and prepare a problem statement to justify how your model could provide value to an organization or to individuals.

Step 2: **Exploratory Data Analysis and Data Cleansing**

Examine your data, and flag out any interesting trends, anomalies, or potential errors. You may need to utilize the below techniques in this step:

- Grouping and Filtering
- Drop Unnecessary Columns
- Missing Value Treatment

Step 3: **Data Wrangling and Transformation**

Wrangle and/or transform the tabular data before feeding it into the Machine Learning portion of your Predictive Model. You may need to utilize the below techniques or apply the following considerations in this step:

- Categorical Data Encoding
- Numerical Data Transformation
- Feature Scaling

**** COMPULSORY TECHNIQUE ****

- Execute use of "Pipeline" from "pyspark.ml" as part of this step where appropriate

Step 4: **Machine Learning Modelling**

State number of rows and columns in your final dataset before building machine learning models, and show a **sample of about 10 rows** before loading into the ML algorithm.

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.

**** IMPORTANT ****

You will be **graded on the use of PySpark**, so usage of **Pandas itself should be avoided as much as possible**, especially if a particular native method or function is already available in PySpark. **Penalties will be imposed in such cases.**

Build a simple machine learning model to predict the resale prices of any given HDB resale transaction.

Step 5: **Model Evaluation and Selection**

Evaluate the model performance. Are you happy with the model performance? If not, please review and tune any of previous steps accordingly.

Justify selection and illustrate comparison between different models before confirming the final model choice.

Step 6: **Report**

To be done as per next section's guidelines.

3. 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, such as screenshots of plots or tables. **You are strongly encouraged to justify/explain choice of treatment or methods used, for each section in the report.**

	Suggested Report Sections & Content Guidelines	**Word Count Guide**
0.	Table of Contents	NA
1.	Problem Statement Formulation • Load and Explore the Data • Understand the Data • Formulate a Value Based Problem Statement	Min: 100 words Max: 200 words
2.	Exploratory Data Analysis and Data Cleansing • Interesting Trends • Anomalies • Potential Errors • Missing Value Treatment	Min: 150 words Max: 250 words
3.	Data Wrangling and Transformation • Categorical Data • Numerical Data • Others	Min: 150 words Max: 250 words

4.	Machine Learning Modelling • Show Count of Rows and Columns • Sample of 10 Rows before Modelling • Build the Predictive Model	Min: 150 words Max: 200 words
5.	Model Evaluation and Selection • Utilize Model Metrics for Evaluation • Compare Models and Decide on Final Model	Min: 150 words Max: 250 words
6.	Summary and Further Improvements • Summarize your findings • Explain the possible further improvements	Min: 100 words Max: 200 words

4. DELIVERABLES

For this assignment, you must submit all the following:

I. A softcopy **Final Report for Section A** (in Microsoft Word / PDF format)

- Submit via "**ASG1 Section A Report Submission**" in POLITEMall
- Save in Microsoft Word/PDF format with the following file naming convention:

  <Student ID>_<Name>_DDP_ASG1_AY2210

  e.g. s2001111A_JohnKhoo_DDP_ASG1_AY2210

II. The completed **"DDP_ASG1_AY2210.ipynb"** Jupyter Notebook File for **Section B**

- Submit via "**ASG1 Section B Submission**" in POLITEMall
- The Jupyter Notebook is to be clearly labelled using markdown cells.
- Save the completed Jupyter Notebook File with the following file naming convention:

  <Student ID>_<Name>_DDP_ASG1_AY2210

  e.g. s2001111A_JohnKhoo_DDP_ASG1_AY2210

III. Link to **Video Recorded Presentation** for **Section B**

- 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.
- Use your **Jupyter Notebook** for your presentation.
- Copy and paste the YouTube video link in final cell of Jupyter Notebook.

All sections must be completed.

Submit the deliverables no later than **Friday 16th Dec 2022, 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 (please refer to Ngee Ann Polytechnic Plagiarism Policy webpage for more information)

5. GRADING CRITERIA

Section A Comprehensiveness (15 marks) Assessed based on: • content of comparison between Hadoop and Apache Spark • rationale and considerations supporting points raised	**Section A Content Layout (15 marks)** Assessed based on: • report style and formatting • visualization and communication effectiveness			**Section A Total (30 marks)**
Section B PySpark Usage and Data Transformation (40 marks) Assessed based on: • PySpark techniques used • no usage of Pandas library • work showing good rationale and considerations of datasets and wrangling and transfomation techniques chosen	**Section B Analysis and Discussion (10 marks)** Assessed based on: • rows and columns of pre-modeling data shown • model comparison, evaluation and selection • showcasing good conclusions from work done • discussion of any trends, anomalies or errors found	**Section B Report Writing (10 marks)** Assessed based on: • effective use of visualizations, appropriate vocabulary, and conciseness • report to be at bottom of Jupyter Notebook, though additional comments throughout the notebook will not be penalized	**Section B Presentation Skills (10 marks)** Assessed based on: • whether presenters show clear understanding of work done • meeting typical video presentation norms (video on, adequate sound level, etc.)	**Section B Total (70 marks)**