اونيۏرسيتي مليسيا ڤهڠ UNIVERSITI MALAYSIA PAHANG PUSAT SAINS MATEMATIK **CENTRE FOR MATHEMATICAL SCIENCES**	**SUBJECT: BSD2333 DATA WRANGLING**		**MARKS:** 15(3%)
	TOPIC: Chapter 3		
	LAB REPORT 2	**DATE:** 12 MAY 2022	

CLO	Description	PLO mapping	Percentage	Marks
CLO1	Acquire data wrangling fundamental concepts and knowledge.	PLO1: Knowledge and Understanding C3: Application	1%	5
CLO2	Apply data wrangling techniques to handle heterogeneous and distributed data.	PLO2: Cognitive Skills and Functional work skills with focus on Numeracy skills C3: Application	1%	5
CLO3	Manipulate data to required format and location for data-driven applications.	PLO3: Functional work skills with focus on Practical, and Digital skills P4: Mechanism	1%	5

The objectives of this Laboratory report are to help you to:

1. Understand the terminologies used in data wrangling.
2. apply data wrangling techniques using Python to solve a data wrangling problem.

Question 1: Data Wrangling Process

Identifying missing values in your dataset is probably one of the most difficult parts of data clean-up, and it takes time to get right. Even if you have a deep understanding of statistics and how missing values might affect your data, it's always a topic to explore cautiously.

1. Explain what is a missing value?
2. Name several missing value detection methods that you know.
3. What are the disadvantages of outliers?

[5 MARKS]
[CO1/PLO1]

Question 2: Wrangling with the Titanic Dataset

In this activity, we will detect and handling missing values in the Titanic Dataset. We will use the concepts we've learned throughout this chapter.

These are the steps that will help you solve this activity:

1. Load the necessary libraries.
2. Read the Titanic dataset.
3. Use the describe command to get the statistical summary of the dataset.
4. Create a DataFrame with only Survived, Age and Fare column.
5. Find the missing values. Check how many NaN values are there.
6. Replace the missing data with a mean value.
7. Find the median and replace the NaN values with the median value.
8. Replace the NaN with the mode value.
9. Decide which one is the best to replace the missing value (mean, median or mode)? Reason.
10. Plot a histogram of Age with a bin size of 20.
11. Plot box plots for Age grouped by Pclass.
12. Find the number of people who are aged between 30 and 50.

[10 MARKS]
[CO2/PLO2 & CO3/PLO3]

اونيۏرسيتي مليسيا ڤهڠ UNIVERSITI MALAYSIA PAHANG PUSAT SAINS MATEMATIK **CENTRE FOR MATHEMATICAL SCIENCES**	**SUBJECT: BSD2333 DATA WRANGLING**		**MARKS:** 15(3%)
	TOPIC: Chapter 3		
	LAB REPORT 2	**DATE:** 12 MAY 2023	

QUESTION 1

CO1: Acquire data wrangling fundamental concepts and knowledge.						
Item Assessed (Cognitive)	**Poor 1**	**Fair 2**	**Good 3**	**Very Good 4**	**Excellent 5**	**Score**
Understand and explain the concept of data wrangling process	Able to identify some of the keywords but fail to give the explanations.	Able to identify some of the keywords but fail to give the explanations.	Able to identify all the keywords but manage to give some of the explanations.	Able to identify all the keywords but manage to give most of the explanations.	Able to identify all the keywords and successfully give the explanations .	
	Total Score					**/5**

QUESTION 2

CO2: Apply data wrangling techniques to handle heterogeneous and distributed data.						
Item Assessed (Cognitive)	**Poor 1**	**Fair 2**	**Good 3**	**Very Good 4**	**Excellent 5**	**Score**
Using analytical, logical or problem solving appropriate to the discipline.	The work has not demonstrated analytical, logical or problem solving understanding appropriate to the discipline.	The work has demonstrated some analytical, logical or problem solving understanding appropriate to the discipline.	The work has demonstrated analytical, logical or problem solving understanding appropriate to the discipline.	The work has demonstrated a thorough analytical, logical or problem solving understanding appropriate to the discipline.	The work has demonstrated a thorough and sophisticated analytical, logical or problem solving understanding appropriate to the discipline.	
	Total Score					**/5**

CO3: Manipulate data to required format and location for data-driven applications.						
Item Assessed (Psychomotor)	**Poor** **1**	**Fair** **2**	**Good** **3**	**Very Good** **4**	**Excellent** **5**	**Score**
Code execution	Code does not work.	Code work but has major flaws.	Code mostly works, and has only minor flaws.	Code works in a way the student intended.	Code is functional and refined with extra features that exceed the requirements.	
	Total Score					**/5**