TRƯỜNG ĐẠI HỌC VĂN LANG
ĐƠN VỊ: KHOA THƯƠNG MẠI

ĐỀ THI/ĐỀ BÀI, RUBRIC VÀ THANG ĐIỂM
THI KẾT THÚC HỌC PHẦN
Học kỳ 2, năm học 2023-2024

I. Thông tin chung

Tên học phần:	KHAI THÁC VÀ PHÂN TÍCH DỮ LIỆU		
Mã học phần:	232_72MISS40233	Số tin chỉ:	3
Mã nhóm lớp học phần:	232_72MISS40233_01		
Hình thức thi: **Dự án/Đồ án/Bài tập lớn/Tiểu luận**	Thời gian làm bài:	**2**	Tuần
☒ Cá nhân	☐ Nhóm		
Quy cách đặt tên file	***Student ID_Student Name_232_72MISS40233_01_Final***		

Giảng viên nộp đề thi, đáp án bao gồm cả **Lần 1 và Lần 2** **trước ngày 15/03/2024**.

1. Formatting Guide

- Assignment total length should not be constrained,
- Individuals submit a soft copy of your finished work at the end of the semester. The soft copy should be submitted through the CTE website of VLU and on the E-learning system.
- This is a group assignment.

2. Sending Assignment Cover Sheet

After approving the assignment cover sheet, answers/rubric, the Head of Department/ the send assignment cover sheet to Trung Tam Khao Thi via email khaothivanlang@gmail.com including Word and Pdf files (compress and set a password for the compressed file) + messaging + naming via tel no. **0918.01.03.09** (Phan Nhất Linh).

II. Intended Course Learning Outcomes Assessed

CLO	CLO Details	Asessment Methods	CLO weight in assessment component (%)	Question No.	Maximum Grade	Matching PLO/PI
(1)	(2)	(3)	(4)	(5)	(6)	(7)
CLO1	Analyze business-related data to provide solutions to improve efficiency for businesses.	Group Project (with Presentation)	30%	1	3	ELO2 (H) ELO3 (S)
CLO2	Apply qualitative and quantitative research methods and tools to analyze, synthesize, and evaluate data and information about the business activities of the enterprise.	Group Project (with Presentation)	30%	1	3	ELO9 (H)
CLO3	Effectively apply teamwork skills, and independent working skills to develop yourself and perform work effectively	Group Project (with Presentation)	20%	1	2	ELO9 (H)

CLO4	Show a professional style, and a sense of responsibility at work, and comply with professional ethics and business ethics.	Group Project (with Presentation)	20%	1	2	ELO10,11 (S) ELO12 (H)

III. Assignment Requirement

❖ Task 1. REGRESSION PROBLEM

Using the dataset **Real_Estate_Data.csv** and fulfill the following requirements (3.5 points)

1) Data Preparation and Initial Examination

- Load and examine the dataset. Describe its structure and any initial observations about missing data or potential issues.
- Handle missing values in **SqFt**, Bedrooms, and **Year_Built** appropriately. Justify your choice of imputation method.

2) Exploratory Data Analysis (EDA)

- Conduct a distribution analysis for continuous variables. Identify any skewness or outliers and discuss potential transformations if needed.
- Analyze the **AreaType** variable in relation to house prices. What insights can you draw?
- Perform a correlation analysis among continuous variables and between each variable and **Price**. Highlight any significant findings.

3) Regression Analysis Preparation

- Discuss how you would prepare **AreaType** for inclusion in a regression model. Implement this preparation.
- Assess multicollinearity among predictors. Would you exclude any variables from the model based on this analysis?

4) Multiple Linear Regression Analysis

- Build a multiple linear regression model using the prepared dataset. Include a brief description of your model specification.
- Interpret the model summary, focusing on the R-squared value, coefficients of predictors, and their statistical significance. What do the results tell you about the factors influencing house prices?

5) Model Evaluation and Conclusion

- Evaluate the overall performance of your regression model. How well does it predict house prices?

- Based on your analysis, what recommendations would you make to a real estate company looking to price their listings competitively?

❖ TASK 2. CLASSIFICATION PROBLEM

Using the dataset **Customer_transaction_record.csv** to fulfil the following requirements (3 points)

1) ***Applying statistical Quartiles and RFM method (1) and K-means (2) to segment customers.***
2) ***Compare the result of the two methods.***
3) ***All the analytical steps and respective results must be reported.***

Instruction for deployment of K-means:

1) Create a SCATTER PLOT.
2) Based on a provided scatter plot, decide the value of K.
3) Select Initial Centroids: Choose k data points from the dataset as the initial centroids.
4) Assign Points to Clusters:

- For each point in the dataset, calculate the distance to each centroid and assign the point to the nearest cluster. You may use the Euclidean distance formula for this purpose.

5) Recalculate Centroids:

- After all points have been assigned to clusters, recalculate the centroids of each cluster. The new centroid is the mean position of all points in the cluster.

6) Iterate the Process:

- Repeat the process of assigning points and recalculating centroids. Continue this iterative process until the cluster assignments no longer change. Document each iteration step, including the positions of centroids and the points assigned to each cluster.

❖ Task 3. DATA VISUALIZATION

Using the dataset **Walmart_retail_data.csv** to fulfill the following requirements (3.5 points)

✓ **Section 1: Dataset and Industry Overview**

1.1 Industry Identification

Task: What industry is highlighted by the dataset?

Example: If the dataset includes sales data, product information, and customer demographics, the industry could be Retail.

1.2 Industry Introduction

Task: Write a one-page introduction about the industry, focusing on key characteristics and its significance.

Example Task: Discuss the Retail industry, emphasizing trends like online vs. in-store sales, the impact of technology, and customer service excellence. Mention the dataset's regional context (e.g., North America) and specific challenges or opportunities.

1.3 Dataset Structure

Task: Describe the dataset's structure, categorizing columns by type.

Example Task: The dataset includes Product_ID (categorical), Sales_Amount (numeric), Transaction_Date (date/time), Store_Location (geographical), and Customer_Age_Group (categorical). Explain the relevance of each to retail analysis.

1.4 Missing Values Analysis

Task: Use Tableau to identify and visualize columns with missing values.
Example Task: Create a bar chart showing the percentage of missing values per column. Discuss how missing Customer_Age_Group data might affect demographic analysis.

1.5 Missing Values Treatment

Task: Discuss strategies for handling missing data in Tableau.
Example Task: Suggest imputing missing Customer_Age_Group based on the mode age group at the same store location. Justify the choice of imputation method.

✓ Section 2: Preparing for a Diverse Tableau Analysis

2.1 Column Selection for Analysis

Task: Justify the selection of specific columns for the analysis.
Example Task: Explain choosing Sales_Amount, Transaction_Date, and Store_Location for analyzing sales trends, highlighting their importance.

2.2 Analysis Objectives

Task: Define your analysis goals, detailing the expected insights and outcomes.
Example Task: Aim to uncover monthly sales trends, top-performing regions, and sales distribution among different age groups, explaining the significance of these insights.

2.3 Variable Mapping for Visual Diversity

Task: Specify the variables and calculated fields in Tableau for each analysis goal.
Example Task: Use Transaction_Date and Sales_Amount for monthly sales trends. Add Store_Location for regional analysis and Customer_Age_Group for demographic distribution.

2.4 Creation of New Variables

Task: If new variables are necessary, list and define them, including the formulas used.
Example Task: Create a Month calculated field from Transaction_Date to aggregate sales by month, detailing the formula and its relevance to the analysis.
Section 3: Tableau Visualization Diversity and Insights

✓ Section 3: Visualization

3.1 Visualization Diversity Requirement

Task: Create at least ten different types of visualizations in Tableau.
Examples:
Bar Chart: Monthly sales to showcase seasonal trends.
Map: Sales by region to highlight geographic performance.
Pie Chart: Sales distribution across demographics.
Line Chart: Sales trends over the year.
Scatter Plot: Age versus purchasing power.
Heatmap: Sales activity by day and time.
Tree Map: Product categories by sales volume.

Bubble Chart: Comparative profit analysis by product.
Box-and-Whisker Plot: Sales variability among stores.
Histogram: Customer age distribution.

3.2 Insights from Diverse Visual Formats

Task: For each visualization type, interpret the insights gained and discuss their implications.

Example Task: Analyze a line chart of monthly sales, identifying a peak in December as indicative of holiday shopping. Discuss how this insight could influence inventory and marketing strategies.

✓ **Section 4: Conclusions and Strategic Recommendations**

4.1 Summary of Observations

Task: Summarize the key findings from your Tableau visualizations.

Example Task: Highlight significant trends such as demographic sales dominance, regional sales disparities, and seasonal buying patterns, emphasizing their implications.

4.2 Strategic Improvement Suggestions

Task: Based on the insights, propose actionable recommendations.

Example Task: Suggest a targeted marketing campaign before the holiday season and stock adjustments based on demographic preferences, justifying these strategies with your analysis findings.

Presentation requirements:

- Present the report in class for 20 minutes for each group;
- Design report using presentation slides which should be submitted individually through the CTE website and on the E-learning system;
- You should use in-text references and a list of all cited sources at the end of the report by applying APA referencing style.

2. Style and Formatting Guide

- The assignment's total length should not be constrained;
- Please submit a soft copy of your finished work at the end of the semester. The soft copy should be submitted individually through the CTE website of VLU and on the E-learning system;

File Naming: *Student ID_Student Name_232_72MISS40233_01_Group Name_Final*;

- It is compulsory to submit the assignment on the due date and in a way requested by the Lecturer.
- This is a group assignment.

3. Grading and Rubric

Criteria	**Weighing (%)**	**Very Good 8-10 pts**	**Good From 6 – under 8 pts**	**Average From 4 to under 6 pts**	**Poor Under 4**

Presentation Time & Report Format	20%	- Perfect timing - Slide Quality: Layout & Visual clear and clean - The group works well, shares and supports each other	- Adequate Timing (± 30 seconds) - Slide Quality: Layout & Visual clear and clean - The group works well, shares and supports each other	- Too short or too long (± 1 minute) - Slide Quality: Layout & Visual averagely clear and clean - The group works well, shares and supports each other	- Finish abruptly (over ± 1 minute) - Slide Quality: Layout & Visual unclear and unclean - The group works well, shares and supports each other
Report Content	30%	- Very well precise, scientific - Use numbers and academic researches for evidence - Sequence information and ideas logically and skillfully; coherent progression throughout	- Mostly precise, scientific, still produce occasional errors - Use numbers and academic researches for evidence, still lack of accuracy - Arrange information and ideas quite logically; clear progression throughout	- Some precise and scientific, may make some errors - Limited use numbers and academic researches for evidence information and ideas relatively logically	- Limited precise and scientific, make noticeable errors - No numbers and academic researches for evidence - Lacks structure and is difficult to follow.
Creativity	20%	- Use creative tools and language, to convey content in a unique understanding of the topic	- Express quite clear and creative arguments	- Express clear and creative arguments averagely	- No arguments

Presentation skills	20%	- Present issues and arguments attractively and persuasively - Very good interaction with audience	- Present clearly but unattractive; arguments are quite persuasive - Good interaction with audience	- Difficult to follow but still able to understand important contents - Inadequate interaction with audiences	- Present unclearly, may not understand - No interaction with audiences
Q&A	10%	- Answer all questions correctly	- Answer all questions, still some errors	- Answer some questions	- Cannot answer

Ho Chi Minh City, 20th March 2023

Internal Verifier **Lecturer**

MSc. Trần Nguyễn Hải Ngân **Ph.D. Lương Thái Hà**