Hands on Data Analytics - Group Project Guidelines

Submission deadlines: November 21, 2022 (Proposal) and December 4, 2022 (Presentation) on iSpace

Submission notes:

- One team member shall submit the project for the for the whole group.
- The proposal submission must be 1 file in .pptx or .pdf with presentation slides (check template)
- The final presentation submission is also 1 file in .pptx or .pdf with presentation slides
- Relevant material that are not presented during your oral presentation shall be included in an Appendix section of the slide deck (for example screenshot of your KNIME workflow, data descriptions, additional analysis results,…). This will be subject to evaluation.
- All submissions will be checked for plagiarism. It is better you do a poor work rather than copy a good work

---

Project Scope

This is a team project. The recommended size for the team is 4 to 6 people (bigger team will not be penalized with respect to smaller teams).

Your goal is to show your understanding and your practical implementation skills of a data analysis project covering the following 3 main area:

1. Import data and create basic exploratory plots
2. Data cleaning, transformation, summary and plotting
3. Train a machine learning model

- **Originality** in presenting and conducting your work **is highly appreciated.**
- Your solution to **difficult steps** in data analysis is **highly evaluated** (for example cleaning very dirty data)

Guidelines for Choosing a Dataset

Available datasets:

- 15 datasets are given in "Group Project" section in iSpace. These datasets have been chosen by our instructors, so their usability and feasibility should be good.
- Alternatively, you may look for one from public dataset hub such as
    a. https://www.kaggle.com/datasets
    b. https://archive-beta.ics.uci.edu/ml/datasets
    c. https://data.sh.gov.cn/view/data-resource/index.html

Please be careful with the size and quality of the datasets. Some of them are too big/small for our project. Some come with poor quality. If your team would like to choose your own dataset, please check with your instructor for approval, so that **the usability of the dataset and the feasibility of the project** can be assured.

Rule of thumb – Choosing **a dataset that you are interested in and curious about** may give you a better chance to succeed in the project.

Guidelines for Choosing the Analysis Question

After selecting a dataset, you should decide **one or a few analysis questions** for your project topic.

- Good questions are asked after you have done exploratory data analysis and basic visualizations. Better questions are asked if you have **domain knowledge** about the subject. For example, if the project involves home loan data, an understanding of the US real estate market, mortgage rate and default trends as well as general consumer sentiment will definitely help you create interesting questions.

- These analysis questions should be **meaningful**. That is, the questions your team try to answer should be **beneficial** to some stakeholders, whether it is yourself, a business organization, or the public.

- The questions shall be **actionable**, meaning stakeholders can take action based on the answer. To find out if they are actionable, you can make one follow up question: "What can the stakeholders do if the answer is x?" where x is a possible, yet random answer. If you can answer that follow up question, it often is a good question.

The list in below provides a few sample questions for your reference. These general questions could be further customized based on your own project and dataset. The list is not exhausted, and you can certainly design your own analysis questions.

d. What types of something (e.g., customers, weather, market, etc.) are we dealing with?
e. What is the probability that something occurs?
f. What are the deciding factors for something to occur?
g. What could happen if some phenomenon happens? (e.g., the beer-diaper example)
…

Recommended Analysis Steps

Before starting the work meet with all team mates. Produce a document organizing your work. For example you could decide to split the work. This document can be included in the presentation slides.

The following steps of the analysis will be subject to evaluation:

Step 1. [Data Import] Import the data in KNIME and understand all different data fields. Identify your modelling goal (data field to predict or main technique to use) and produce a basic data summary plot

Step 2. [Data Preparation] Clean and filter the data and produce new data features (feature engineering) as you think they are useful for your modelling goal.

Step 3. [Modelling] Apply any machine learning model (e.g. regression, classification, clustering,..) that is most appropriate for your modelling goal.

Step 3. [Result] Create plots, tables and an interpretation, summarizing your results.

Project Presentation Guidelines

- Your group presentation shall last at most 12 minutes with an additional 5 minutes for discussion.
- We recommend that you prepare about 10 slides covering the following points
  - Business problem and data description (Introduction)
  - Data preparation you applied
  - Modelling techniques you applied
  - Result / Conclusion from your analysis
- We recommend you use plenty of plots and images during your presentation.
- All team members must participate in the presentation. Team members who fail to present will be highly penalized and could receive a 0 score.
- It is not recommended to read a script during the presentation. You can use a note with keywords.