Hong Kong Baptist University Department of Computer Science

COMP3115 Exploratory Data Analysis and Visualization Semester 2, 2022-23

Group Project

Select ONE from the following two topics for your project:

Topic 1: Open Covid-19 Data Analytics

Objective:

This project is designed for students to apply what they have learned in the lectures and lab sessions to real-world data for exploratory data analysis and visualization. In particular, students are expected to identify meaningful data analysis tasks, acquire the relevant datasets, select the machine learning algorithms for the analysis, evaluate the performance, and present the findings and insights gained using appropriate visualization methods.

Project Details:

The theme for this year is related to Covid-19 which has been affecting many aspects of our life. Related data sources have been made available online and open for people to make further use of them. Such open data can enable us to carry out analysis to gain better understanding of the pandemic. In this project, the suggested data sources are from:

https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent

You are expected to form groups (**five** students) to work on the project. The project should include the following steps:

1) *Identify the data analysis tasks*

- You are expected to identify at least **FIVE** data analysis tasks. While you are free to identify the tasks of your own choice, the following two are provided for your reference. It is up to your group to adopt them or not. In general, the more diverse the four tasks are, the higher mark will be gained. Also, if the four diverse tasks together can further tell a coherent story, even better with a higher mark!
- Example 1 (Classification): Given the number of confirmed cases for a period of time, predict if the number of confirmed cases will increase or decrease in the following days.
- Example 2 (Clustering): Group the weeks with a similar pattern of mode of detection of

cases.

2) *Identify the data sources relevant to the tasks*

- For Example 1, you can retrieve the number of cases over a reasonable period of time from the data.gov.hk website. Also, study the data dictionary if provided and browse through the data to have better ideas about the features included. If you think that the features are not sufficient, see if there are some additional features you can find on the website (or other sources) to be included. Notice that you need to align the features from multiple data sources by the reference dates of the data items.
- For Example 2, you can retrieve the mode of detection of cases again from the website.

3) *Conduct simple data analysis and visualization*

- Tabulate and plot the acquired data and their statistics to gain better understanding of the data for your four tasks (e.g., plot of cases over time, cases per district, etc.).

4) *Apply machine learning algorithms for the analysis*

- Apply appropriate machine learning (classification, regression, clustering) algorithms to the acquired data and visualize the results. If the number of features is large, dimension reduction techniques (e.g., PCA) can be adopted before applying the machine learning algorithms.
- You are encouraged to use different algorithms, try different settings (parameter settings, experiment settings), compare their performance, and discuss your findings (including the comparison of different algorithms and the insights about the pandemic gained from the results).

5) *Discussion*

- Discuss the limitations of the studies and suggest ways to address the limitations as future work.

Topic 2: Self-proposed project

You can also propose your own projects. However, the topic of the project should be closely related to exploratory data analysis and visualization. **The requirements and workflow should be the same as Topic 1.** If you prefer to propose your own projects, please submit a one-page project proposal to me (ericluzhang@comp.hkbu.edu.hk) and Kenny (kennycheng@comp.hkbu.edu.hk) before **5:00pm, 10 March, 2023 (Friday).** In this one-page proposal, it should contain: (1) the motivation and objective of the project; (2) data sources and formats; (3) a list of proposed tasks (five tasks); (4) the expected outputs of the project. We will take a look at your proposal and reply to if your proposed project is suitable for the course project.

Project Implementation:

You can select to use python or other off-the-shelf tools to implement the project, but the grades will be marked separately.

Group information submission:

Please submit your group members on Moodle before **5:00pm, 3 March, 2023 (Friday)**, even if your group members are less than five. We can help reorganize the groups.

Project Submission:

Each group is required to submit (one set per group is sufficient):

i) *Source file (zipped)*: One zipped file with the filename "COMP3115_GpX_source.zip" with either **1. jupyter notebook file and data file** (using python) or **2. screenshots from the tools with necessary descriptions in pdf format and data file** (using off-the-shelf tools). A file should be submitted per group with different sections corresponding to the different tasks.

ii) *Project report*: A project report (pdf format) with the filename

"COMP3115_GpX_report.pdf" documenting the details of all the steps. Include in the report a table showing the contribution of each member to different parts of the project. The maximum number of pages is 20 (concise and precise writing is expected; figures and illustrations are included in the page limit; title page and references are not included in the page limit; 1.5 line spacing; 1 inch for all margins; use new times roman point 12 font).

iii) *Presentation slides*: The file with the project presentation slides (pdf format) with the filename "COMP3115_GpX_slides.pdf".

Important Dates:

April 13, 19, 20	Group Project Presentation - Each group will be allocated 15 minutes for project presentation (10 min.) and Q&A (5 min.). All members should present.
April 25 (17:00)	Submission • Source file • Project report • Presentation slides

Important Notes

All the assignments and the project report, if not otherwise specified, have to be written in your own words instead of copy-and-paste from the Web or AI-based tools, and submitted via HKBU Moodle with plagiarism checking. Plagiarism cases once caught will receive a heavy penalty.