*The Hong Kong Polytechnic University*

**COMP 1433: Introduction to Data Analytics**

**Final Project Specifications**

**Important Notes (Read before you get started!).**

1. The project will account for 45% of the final grade.
2. The project is to be completed in groups, i.e., one group only needs to submit one project report (and one attachment with R scripts if needed).
3. Paper size: A4; font: 12-pt Times New Roman; language: English; margin: 2.5cm (each of the four sides); line spacing: 1.0 (single line). [[1]](#footnote-1)
4. Page limit: at most 8 pages (*no lower bound*).
5. Submission deadline: **23:59 May 14, 2021** (Friday)
6. No late submission will be allowed.
7. There are three options to be introduced in below and each group can pick up any one of them. No matter what option you go for, a project report is needed to be submitted. For option 2 and 3, in addition to the report, please also submit an attached zipped file with the R scripts used to do data analysis and model implementation (so we can review the contributions from the group members working on programming).
8. The project report should be submitted via *Blackboard* (https://learn.polyu.edu.hk) and the entry is *Assessments/Project/Report*. The attachment (R scripts for option 2 and 3) should be zipped and submit to *Assessments/Project/Attachment.*
9. Plagiarism is not allowed. Any suspected cases will be reported to PolyU and processed according to the university's regulations.
10. There is no guarantee that everyone in a group will get the same grades. To encourage everyone to contribute more to the teamwork, we’ll set the weights to reflect the individual’s contributions. We’ll set up a questionnaire (will be available at Assessments/Project/Inter-Annotations on Blackboard) for each student to evaluate other members’ contributions in the group. The individual contribution weight (denoted as ω) will be determined by the median of other members’ score (in 100). If the group’s score is γ, the final score of an individual shall be ω100⋅γ. For example, if student S gets 90 for group project, and other members assign 100, 95, and 70 for S. The final score for S is 90⋅95/100=85.5. If one didn’t submit the questionnaire, we’ll assume that they gave 100 (full mark) for every other members as the individual contribution weights

**Detailed Requirements.**

* ***Option 1: Literary Review***

This option is for students who are not a big fan of programming yet like to learn new things for Artificial Intelligence Concepts. In the literary review project, each group should first pick up a topic related with this course, e.g., *Comparison of Diverse Searching Algorithms*, *Development of Statistical Models*, *PageRank Model and its Expansions*, etc.

Then, the group should read at least 5 research papers centring around the topic you select. The papers should have at least 100 citations on Google Scholar and were published in a year after 1980. Here are some suggested conferences and journals where you can find high quality papers:

* **Data Mining**: SIGKDD (<https://www.kdd.org/>), WWW (<https://www.iw3c2.org/>), TKDE (<https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=69>)
* **Information Retrieval**: SIGIR (<https://sigir.org/>), CIKM (<https://dl.acm.org/conference/cikm>), TOIS (<https://dl.acm.org/journal/tois>).
* **Machine Learning**:NIPS[[2]](#footnote-2)(<https://nips.cc/>), ICML(<https://icml.cc/>), PAMI (<https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=34>).
* **Artificial Intelligence**: AAAI (<https://www.aaai.org/>), IJCAI (<https://www.ijcai.org/>), TIST (<https://dl.acm.org/journal/tist>)

In the proposal presentation, you’ll have a chance to discuss the topic you will explore and the papers you want to read. We’ll examine whether the papers and the titles are expected and send feedback to you via proposal comments. You may decide whether you need to change the papers or update the topics by the comments received.

Afterwards, the team work together to write a report that summarize the papers and present a discussion of your thoughts. Your report is suggested to be organized in the following way (you may add other parts to make the story goes more smoothly):

* **Motivation**. The reason why you select the topic, e.g., why the topic is useful and who can be benefited from the task?
* **Background**. The history and development of the related area, e.g., what previous research did on the task. Such information can be found on related work or background study of the research paper.
* **Description.** Introduce the methods proposed by the papers in an easy-to-understand way (which means that more description (with figures) and less formulas). Summarize the contributions made by the papers you read. The goal here is to make sure that people who have not read the papers are able to capture the ideas quickly.
* **Data.** A brief description about what the data is used for experiment in the papers you read.
* **Discussions.** Your thoughts after reading the papers. You can share the inspirations gathered from the papers or the suggestions you propose to further improve the work.
* **Reference.** List of the papers you surveyed and cited.
* ***Option 2: Data Analysis for House Prices***

This option is for students who are a big fun of programming for data analysis. The topic is inspired by Kaggle House Prices competition (<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>). Here is the start of the story --- *Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.*

*With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.*



In this project, you will be able to figure out what is probably the final price of each home. You will be given the information of each aspects of residential homes in Ames, Iowa (in the attached file *train.csv*), such as fence quality, pool quality, garage condition, year garage was built, and will have the chance to use the tools (e.g., statistics, regression, graphs) learned in COMP 1433 to analyze the key factors (features) resulting in prices of the home. Afterwards, please learn how to implement Regression model (Linear Regression, Logistic Regression, etc.) and the use the features (pool quality for example) you consider important to produce a Regression model that is able to predict the prices (in *test.csv*) of the homes. At last, you can submit your results to Kaggle to get the results (Please go to the Kaggle page <https://www.kaggle.com/c/titanic> to see how to submit the predictions and refer to the evaluation metrics which is Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally). You will be allowed to have 10 submissions per day on Kaggle.

The training data is provided on Blackboard on Assessments/Project/option2\_data.

In the report to be submitted, the organization is suggested to be in this way:

* **Motivation**. The reason why we should understand the Price data (from the perspectives of history, social science, and so forth).
* **Description.** Describe the methods (or tools) you used to analyse the features and why they can be helpful. Also, give a brief introduction of Regression models (in your own words).
* **Implementation.** How you implement the Regression models (describing both variables and functions used in the R scripts).
* **Data.** Present the key statistics in the data you are working on (both train.csv and test.csv), such as the average value, range, distributions, etc.
* **Results and Observations.** Show your analysis results (in figures, tables, or numbers) and list the observations drawn from the results. Here please also provide the screenshot of your team profile and the scores you got from Kaggle.
* **Discussions.** Your thoughts and opinions after analysing the data (can be from different perspectives of history, social science, etc.).

In addition to the report, please also submit a zipped file with the R scripts used to analyse the data and implement the Regression models. Please make sure that the codes are consistent with the implementation parts in the report and commented well to allow reviewers to capture the key idea.

* ***Option 3: DIY your own task***

This option is for those who want to challenge themselves and are not interested in either option 1 or 2. So, you can have the choice to DIY your own task, including problem definition, data collection, methodology design, results analysis, etc. The task should be related with data analytics, machine learning, and other contents learned in COMP1433. For option 3, you are required to submit a report covering the following parts:

* **Motivation**. The reason why the task you pick up is important and who can be benefited from the task you tackle.
* **Background**. The history and development of the related area, e.g., what previous research did on the task. Present a short literary review with brief description, which is different from option 1 (where you need a detailed introduction).
* **Description.** Firstdefine task you are exploring. Then, describe the methods (or tools) you used to analyse the data and explain why they can be helpful.
* **Implementation.** How you implement the data analysis tools (describing both variables and functions used in the R scripts).
* **Data.** Describe the data you are using for analysis and the way you collect them. Present the key statistics in the data you are working on, such as the average value, value range, distributions, etc.
* **Results and Observations.** Show your analysis results (in figures, tables, or numbers) and list the observations drawn from the results.
* **Discussions.** Your thoughts and opinions after analysing the data (can be from different perspectives such as public health, business, etc.

In addition to the report, please also submit a zipped file with the R scripts used to analyse the data. Please make sure that the codes are consistent with the implementation parts in the report and commented well to allow reviewers to capture the key idea.

1. Slightly changing the fonts for titles, footnotes, figure and table captions as well as tuning paragraph spacings to allow better display are acceptable. [↑](#footnote-ref-1)
2. Changed to NeurIPS recently. [↑](#footnote-ref-2)