**Final Projects**

Students will work on a month-long data science project. The goal of the project is to go through the complete data science process to answer questions you have about a topic of your own choosing. You will acquire the data, design your visualizations, run statistical analyses, communicate results, and write technical reports.

**Project Team**

Students will work in teams of 2-4 and will work closely with other classmates on this project. In general, we do not anticipate that the grades for each group member will be different. However, we reserve the right to assign different grades to each group member based on peer assessments and class attendance.

**Project Milestones**

There are a few milestones for the final project. **It is critical to note that no extensions will be given for any of the project due dates for any reason, except for COVID-19 related emergencies or other unforeseen emergencies.** Projects submitted after the final due date will not be graded. Students who anticipate any issues should send an email to me at least one week in advance.

| **Date** | **Description** |
| --- | --- |
|  |  |
| December 8 by 11:59pm (Bangladesh Time UTC+6) | Form a team and submit a project description in the online spreadsheet (<https://docs.qq.com/sheet/DZVprclhYTEtacVNN>). |
| December 22 | Project mid-term review due (submit your progress report in PowerPoint slides to your class president: AHMED ISTAK) |
| January 7 by 11:59pm | Project files submission due (code, data, technical report) |
| January 8 by 11:59pm | Peer assessment due |

**Deliverables**

There are several deliverables for your project that will be graded individually to make up your final project score.

**Team Registration and Mid-Term Review**

Students start by filling out an online form to define your teams and project proposal. The subtasks each member will be responsible for can also be initially stated. This form should be filled out by 11:59pm (UTC+6) on December 8, 2021. The title may be changed at a later date (within 5 days) if needed. Each team will need to submit a slide (~5 pages) to report your progress before December 22. The Kaggle website is an excellent reference for you to formulate your proposal and acquire a data set.

**Technical Reports**

An important part of the project is the technical report (TR). It should detail the steps taken in developing a solution(s), including how students collected the data, alternative solutions tried, statistical and/or machine learning methods used, experimental results, and insights. Equally important to the final results is how the team got there! The technical reports are the place you describe and document the space of possibilities explored at each step of the project. We strongly advise you to include many visualizations.

The TR file should include the following topics. Depending on the project type, the amount of discussion devoted to each will vary:

* **Overview and Motivation**: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.
* **Related Work**: Anything that inspired you, such as a paper, a web site, or something we discussed in class.
* **Initial Questions**: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?
* **Data**: Source, scraping method, cleanup, etc.
* **Methods and Results**: What exploratory analysis did you perform? What visualizations did you use to look at your data in different ways? What are the different statistical/machine learning methods you considered? Compare the methods you choose and the results obtained by these methods, justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?
* **Discussion and Future Work**: What did you learn about the data? How did you answer the questions? How can you justify your answers? What future work/improvement would be expected?

**Code and Datasets**

You need to submit all your code and datasets in order to reproduce all the methods and results you have addressed in the technical report. We expect you to write high-quality and readable Python code. You should strive for doing things the right way and think about aspects such as reproducibility, efficiency, cleaning data, etc. **We also expect you to document your code.**

**Peer Assessment**

It is important to provide positive feedback to people who truly worked hard for the good of the team and to also make suggestions to those you perceived not to be working as effectively on team tasks. We ask you to provide an honest assessment of the contributions of the members of your team, including yourself. The feedback you provide should reflect your judgment of each team member:

* **Preparation**: were they prepared during team meetings?
* **Contribution**: did they contribute productively to the team discussion and work?
* **Respect for others’ ideas**: did they encourage others to contribute their ideas?
* **Flexibility**: were they flexible when disagreements occurred?

Your teammates’ assessment of your contributions and the accuracy of your self-assessment will be considered as part of your overall project score. The peer assessment is due January 8 by 11:59pm (UTC+6). For instructions on how to submit, please see **Submission Instructions** below.

**Submission Instructions**

**How to submit your project files**

1. Create a GitHub repository which should include the data used for the final project and the technical report. If the data are too big to fit in the repository, make the data accessible somewhere online (Baidu drive, Google drive, OneDrive, downloadable link, etc). If we cannot access your work or links because these directions are not followed correctly, we will not grade your work.
2. Make sure you have included the technical report in the repository (e.g., in a .docx format) and the names of all group members are inside the technical report at the top.
3. Email me at [whpsdut@163.com](mailto:whpsdut@163.com) the instructions on where to access the data and the location of your GitHub repository.
4. You should also send your code, data and technical report in a ZIP file to me ([whpsdut@163.com](mailto:whpsdut@163.com)) as a backup by January 7 11:59pm (UTC+6). If the data are too big, attach the link in the email.

**How to submit the Peer Assessment**

Each individual team member needs to fill out an online form (which will be released later) for the peer evaluation. Your individual project score will take into account your self and peer assessment.

**Grading**

The final project is graded in two parts:

1. Final Project Part (worth 90% of total grade).
   * 20% of the Final Project Part will be based on the Team Registration, Project Proposal, and Mid-Term Review.
   * 40% of the Final Project Part will be based on your technical report. This includes the complexity and level of difficulty of your project, the quality of your data analysis, the quality of your storytelling aspects, completeness and overall functionality of your analysis.
   * 30% of the Final Project Part will be based on the quality and readability of your code and the code annotation.
2. In-class Performance (worth 10% of total grade).

Your individual project score will also be determined by your peer evaluations.

**Example Final Projects**

Here are some examples of final projects you could refer to.

1. **Predicting Hubway bike/dock availability** ([Website](http://cs109hubway.github.io/classp/), [Screencast](https://www.youtube.com/watch?v=2wK8jpNMjXI&feature=youtu.be))
2. **Across the Bay 10K Race** ([Website](https://sites.google.com/site/atb10kbridgerace/home), [Screencast](https://sites.google.com/site/atb10kbridgerace/home/screencast))
3. **Predicting AirBnb Success** ([Website](http://hamelsmu.github.io/AirbnbScrape/), [Screencast](https://www.youtube.com/watch?v=raGjUj5qArc))