

Course Project

Overview

Analytics is a topic that is best learned by applying the methods to real-world data and problems—hence, analytics in *action*. In this project, you and your team will select a real-world problem to investigate using the techniques learned in lecture and lab. You are expected to: i) identify a problem or application area where analytics can be applied, ii) obtain data and conduct exploratory data analysis (EDA), iii) apply an analytics method or combination of methods to solve your problem/answer your question, iv) analyze the results, and v) present findings and conclusions.

The project consists of the following deliverables:

1. **Proposal (10%):** The proposal presents a problem or question or hypothesis that you wish to explore using analytics. It should also outline the data sources and proposed methods to be used. Projects can deviate from the original proposal, but major changes should be cleared with the teaching team.
Length: One page.
2. **Preliminary report (15%):** The preliminary report summarizes progress to date, including data obtained, initial findings from EDA, initial results, changes to the proposal, and a plan to complete project.
Length: Three pages.
3. **Abstract (5%):** The abstract provides a succinct summary of the entire project, outlining the problem statement, data, methods, results, and conclusions. It will be shared with the other teams and help decide who presents their project during the final week of class.
Length: Half a page.
4. **Presentation (25%):** The presentation summarizes the entire project including problem statement, data, methods, results, and conclusions. All team members are expected to speak.
Length: Five minutes (excluding Q&A).
5. **Final report (45%):** The final report should be a comprehensive write-up of the entire project, including the motivation, a clear problem statement, how and from where data was collected, specific analytics methods applied, results from the analyses, and conclusions and future directions.
Length: Eight pages (not including appendices).

Project Coach

Each team will be assigned one TA as project coach following the proposal submission. This TA will be your primary point of contact for the project and will be invested in helping you succeed.

Data Sources

Each team is expected to identify an appropriate data source and gather the data. The project scope will largely be driven by the data you find and use. Multiple teams may use the same data, but the question must be different. Some possible datasets and associated real-world applications/problems are linked below, but we encourage you to be creative and look into other data sources as well.

- **Toronto Open Data Catalogue:** Contains various datasets on everything from ambulance locations to traffic cameras.
- **HetRec:** Hosted by the University of Minnesota, contains several open datasets for recommender engines, such as Delicious, Last.FM, MovieLens, and IMDB.
- **SQuAD:** Stanford Question Answering Dataset, contains parsed Wikipedia articles and crowdsourced questions for trivia bots and personal assistants.
- **Million Song Dataset:** hosted by Columbia University, contains several open datasets for music information analytics.
- Search for datasets on GitHub, Kaggle, Reddit /r/opendata, and elsewhere!
- If you have an interesting project idea and are missing some crucial data, you could collect your own (for example with surveys, scraping websites, or manual collection).

If you are having a hard time finding a dataset for your project, please contact the TAs before submitting the proposal (or your project coach once assigned).

Proposal

The proposal should focus on defining the topic and scope of the project. It is an opportunity for you to identify an interesting problem and methods to solve it, from data collection all the way to model implementation. The proposal should be no more than **one single-spaced page** and should include the following sections:

- **Background:** Brief background on the domain of your problem.
- **Problem statement:** A clear problem statement identifying the question(s) you will answer.
- **Data:** A description of the sources and datasets that you plan to use, including key variables.

- **Methods:** An outline of the analytics methods and models that you plan to implement in order to answer the question.

The teaching team will review your proposal to make sure that you have selected an appropriate problem and to suggest avenues to explore for data collection and model implementation.

Preliminary report

The preliminary report is a key milestone towards the completion of the project. It is critical for making sure the project is going in the right direction. Putting in a significant early effort to identify challenges and/or promising directions will pay off in the long run. The preliminary report should be no more than **three single-spaced pages** and should include the following sections:

- **Problem statement:** A revised (if necessary) problem statement.
- **Data:** A revised (if necessary) description of data.
- **Methods:** A revised (if necessary) description of the methods and models.
- **Initial findings:** Results from relevant exploratory data analyses, including summaries of key variables, data missingness, etc. Also include initial results from applying the chosen analytics method(s) to the data. The initial findings may be useful in revising the project scope. Make effective and selective use of figures and tables.
- **Completion plan:** Identify the remaining steps to complete the project, including additional data to be gathered, other methods to try, etc.

Abstract

The abstract provides a succinct summary of the entire project, outlining the problem statement, data, methods, results, and conclusions. A reader should get all the relevant details about your project by reading the abstract. All abstracts will be posted on the course website and students will vote for which project they would like to see presented in the final week of classes. The abstract should be a structured abstract with the following sections: Problem statement, Data, Methods, Results, Conclusions. It should be no more than **half a single-spaced page**. At this point, the project should be very close to completion, so the presentation and final report should not deviate much from the submitted abstract.

Presentation

The presentation should summarize the entire project, following the sections from the abstract. Use the presentation to convince the teaching team that you chose

an interesting problem and that your results are meaningful. Be energetic and creative! To ensure crisp presentations and strict adherence to the **five-minute time limit**, we will be using a shortened **Pecha Kucha** format. That is, slides must auto-advance every 20 seconds. Presentations will thus be exactly 15 slides long, excluding the title slide. To be successful in a Pecha Kucha format, teams must invest substantial effort in designing slides, graphics and content, and practicing the delivery of the presentation. Poorly rehearsed Pecha Kecha presentations are obvious.

Presentations will take place during lab time in rooms to be announced. Scheduling details will be provided later. All team members are expected to participate in the presentation. There will be a short Q&A with the teaching team following each presentation.

Final report

The final report provides a comprehensive summary of the project. It should be no more than **eight single-spaced pages (excluding appendices)** and should include the following sections: Introduction, Data, Methods, Results, Discussion, Conclusions and Future Directions, and References. You can also include an Appendix that includes additional technical details, supplementary data analysis, code, sensitivity analyses, etc. Only include material that is essential.