STAT 201

Statistical Inference for Data Science

Group Project

As in DSCI 100, students will work in groups to complete a Data Science project from the beginning (downloading data from the web) to the end (communicating their methods and conclusions in an electronic report). However, this time, your project will focus on statistical inference. Remember, a fundamental theme in this course is that, in addition to drawing conclusions (such as computing an estimate), it's critical to communicate the evidence and uncertainty associated with the conclusion. The latter is the main focus of this project and will manifest in an electronic report.

The electronic report will be a Jupyter notebook in which the code cells will download a dataset from the web, reproducibly and sensibly wrangle and clean, summarize and visualize the data, as well as appropriately answer an inferential question. Throughout the document, markdown cells will be used to communicate the question asked, methods used, and the conclusion reached.

For this project, you will need to formulate and answer an inferential question about a dataset of your choice. We list some suggested data sets below; however, we encourage you to use another data set that interests you. If you are unsure whether your dataset is adequate, please reach out to a member of the teaching team.

Deliverable 1: Team Contract

A group contract is a document to help you formalize the expectations you have for your group members and what they can expect of you. It will help you think about what you need from each other to work effectively as a team! You will create and agree on this contract as a team. **Each member should "sign" (you can just type out your name) at the bottom of the submission**. At a minimum, your group contract must address the following:

Goals

- What are our team goals for this project?
- What do we want to accomplish?
- What skills do we want to develop or refine?

Expectations

What do we expect of one another regarding attendance at meetings, participation, frequency of communication, quality of work, etc.? What are our internal deadlines? (Warning: if working on separate parts, do not aim to put all the parts together on the last day – it takes time to integrate multiple parts.)

Policies & Procedures

What rules can we agree on to help us meet our goals and expectations?

Consequences:

How will we address non-performance regarding these goals, expectations, policies and procedures?

Deliverable 2: Project Proposal

Each group is expected to prepare a written proposal within 500 words (about 1 page) that identifies the dataset they plan to work on, as well as the question they would like to answer using that dataset for their group project. The proposal should be done in a Jupyter notebook, and then submitted both as an `.html` file (`File` → `Download As` → `HTML`) and an `.ipynb` file that is reproducible (i.e. works and runs without any additional files).

Only one member of your team needs to submit. You must submit **two files**:

- the source Jupyter notebook (`.ipynb` file)
- the rendered final document (`.html` file)

Each proposal should include the following sections:

- Title
- Introduction
- Preliminary Results

- Methods: Plan
- References

Introduction

Begin by providing some relevant background information on the topic so that someone unfamiliar with it will be prepared to understand the rest of your proposal.

Clearly state the question you will try to answer with your project. Your question should involve one or more random variables of interest, spread across two or more categories that are interesting to compare. For example, you could consider the annual maxima river flow at two different locations along a river, or perhaps gender diversity at different universities. Of the response variable, identify one location parameter (mean, median, quantile, etc.) and one scale parameter (standard deviation, inter-quartile range, etc.) that would be useful in answering your question. Justify your choices.

Identify and describe the dataset that will be used to answer the question. Remember, this dataset is allowed to contain more variables than you need – feel free to drop them!

Also, be sure to frame your question/objectives in terms of what is already known in the literature. Be sure to include at least two scientific publications that can help frame your study (you will need to include these in the References section). We have no specific citation style requirements, but be consistent.

Preliminary Results

In this section, you will:

- Demonstrate that the dataset can be read from the web into R.
- Clean and wrangle your data into a tidy format.
- Plot the relevant raw data, tailoring your plot in a way that addresses your question.
- Compute estimates of the parameter you identified across your groups. Present this in a table. If relevant, include these estimates in your plot.

Be sure to not print output that takes up a lot of screen space.

Methods: Plan

The previous sections will carry over to your final report (you'll be allowed to improve them based on feedback you get). Begin this *Methods* section with a brief description of "the good things" about this report – specifically, in what ways is this report trustworthy?

Continue by explaining why the plot(s) and estimates that you produced are not enough to give to a stakeholder, and what you should provide in addition to address this gap. Make sure your plans include at least one hypothesis test and one confidence interval. If possible, compare both the bootstrapping and asymptotics methods.

Finish this section by reflecting on how your final report might play out:

- What do you expect to find?
- What impact could such findings have?
- What future questions could this lead to?

References

At least two citations of literature relevant to the project. The citation format is your choice – just be consistent. Make sure to cite the source of your data as well.

Deliverable 3: Peer Review

For this peer review, you will be **individually providing feedback to another group's proposal. You will be sent an email with the proposal** to review.

We deliberately set up the review so that each member of your group is assigned to review a *different* group's proposal. This allows your group to collectively see a larger variety of proposals.

Instructions

Bundle all of your feedback into a single document of your choice, and submit that to canvas. There is no page limit. The teaching team will deliver the feedback to your reviewee.

Your review should contain the following elements:

- As you read through the proposal, point out anything that you think is confusing, or is not communicated effectively. When possible, provide suggestions for improvement. If everything looks good to you, say why it looks good.
- What part of the proposal is the most effective, and why?
- What part of the proposal is the least effective, and why? Provide a suggestion for improvement.
- Provide feedback on English, spelling, and grammar (if applicable).

The rubric can be found on canvas. In short, *mechanics* (10%) evaluates the composition

your English, spelling, and grammar.

Deliverable 4: Final Report

Each group will create a final electronic report (max 2000 written words, not including citations) using Jupyter to communicate the question asked, the analysis performed and the conclusion reached.

Only one member of your team needs to submit. You must submit **two files**:

- the source Jupyter notebook (`.ipynb` file)
- the rendered final document (`.html` file)

Each report should include the following sections:

- Title
- Introduction
- Methods and Results
- Discussion
- References

Introduction

The instructions for this section are the same in your proposal. Just be sure to improve this section by incorporating feedback, and changing things based on your own improved understanding of the project (now that more time has passed since the proposal).

Methods and Results

Here is where you'll include your work from the "Preliminary Results" in your proposal, along with the additional results you planned to conduct, as indicated in the "Methods: Plan" section of your proposal. Be sure to incorporate feedback from the teaching team and your peers (as relevant), or make any improvements based on your own improved understanding of the project (now that more time has passed since the proposal).

Specifically, in addition to what is requested in the "Preliminary Results" section of the proposal, we are looking for the following components:

- Describe in written English the methods you used to perform your analysis from beginning to end that narrates the code the does the analysis.
    - Make sure to interpret the results you obtain. It's not enough to just state what a 90% confidence interval is, for example.
- Ensure your tables and/or figures are labeled with a figure/table number.
- Do you think one of bootstrapping or asymptotics is more appropriate than the other? Why or why not? Explain why you think both methods gave you similar/different results, and whether you think one is more trustworthy than the other.

Discussion

In this section, you'll interpret the results you obtained in the previous section with respect to the main question/goal of your project.

- Summarize what you found, and the implications/impact of your findings.
- If relevant, discuss whether your results were what you expected to find.
- Discuss future questions/research this study could lead to.

References

The same instructions for your proposal also applies here. You only need to make changes if necessary (e.g., if feedback indicates so).

Deliverable 5: Team Evaluation

Evaluate each member of your group (including yourself) in terms of how they/you participated, prepared, helped the group excel, and was a team player.

Click the Teammate Evaluation Template link in the Canvas home page to access the teamwork document. Fill out the jupyter notebook, then download an html rendering of the completed notebook by going to `File` → `Download As` → `HTML`. Finally, submit the rendered HTML document here.

- This is **not** a group submission. Every member of the group must complete and submit this individually.
- This is **not** a standard worksheet/tutorial. You **must** download off our server and submit the html to Canvas.

Data

- UCI Machine Learning Repository has hundreds of datasets
- City of Vancouver data: https://opendata.vancouver.ca/pages/home/ (or other cities, like Ottawa, Toronto, etc.)
- Vancouver crime data: https://geodash.vpd.ca/opendata/
- BC data: https://data.gov.bc.ca/
- Water survey of Canada: https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey.html
- Climate and Weather Data: https://climate.weather.gc.ca/
- Sports data
- Maybe do your own experiment or observational study, if you have time – or maybe you already have the data. For example, do you track your daily step count? If so, you could come up with some criteria to determine whether you have been satisfying the goal, and use your daily step counts. You could see how your performance holds up depending on the day of week (or maybe whether or not it's a weekday), or perhaps across different group members.

Attribution

These instructions are a modified version of the DSCI 100 project at UBC