# Philosophy

The research project is meant to give you a chance to practice and learn some new data science skills.

# Overview

The term project for the course will require you and your group to write a research paper related to the theme for the term. The theme will be announced in class shortly after the start of the semester.

The exact research question(s) are up to you but the question must be reasonably related to the theme this semester.

In order to ensure that you are on track and to provide feedback on your idea, we will ask you to complete a research proposal to explain your research questions and your proposed approach.

# Format

Use the ACM style to cite references.

We will ask that your submission is in PDF form with a 12 point font. Single space is fine (as is double space) but please do use headers and/or some mechanism to make it clear which parts of the assignment the various portions of your submission are meant to address.

# What to Include

Here are some questions that your submission must answer:

1. What are your research questions?

   Generally speaking, we wish to see a research question that is investigated in different ways (one for each team member). One of your research questions must involve an aggregate operation of some time whether it is an average, sum, division, subtraction etc.

   Part of choosing a good research question is to scope it out correctly meaning you need to ask a question that is specific and can actually be answered. The number of research

questions you need to submit will depend on the scope of the research questions and how many members are on your team.

This is a **very** iterative process so be sure to start early and ask questions!

When designing a research question, be specific about what it is that you want to find (see the Example Research Question section later in this document). Consider what values you have access to examine, and what measures you can derive from the examined values. Then, you will need to determine whether those measures can actually help you answer your question.

There are many times when you have a question but there is no way to get the data you need to answer it. In those cases, you will need to iterate over your question again to see how you can tweak it.

2. How will you evaluate your data? How do you know that this is a sound evaluation metric? For example, using correlation to determine causation is not an appropriate thing to do. Provide evidence that your evaluation methodology is scientifically sound.

3. Why do you ask this? What is interesting about this question?

   In the final research paper, we will be asking you to talk about related work and what others have done. You **don't** need to come up with something completely novel and never asked. There's value in replication. You still have to do your own work (e.g,. If you need to write SQL statements to pull an answer from your database you need to write that yourself and you cannot copy something someone else has already done).

4. What data will you get? Please provide links to the location of the datasets. We will ask you to get data from at least two different sources.

5. How do you know that the data is trustworthy? Keep in mind that just because data is from an official department, it does not mean that it is unbiased. You often need to cross reference and/or look up details about how datasets have been compiled in order to adequately address questions around data validity and potential biases.

6. How will you combine your datasets?

7. Don't forget your AI tool use declaration!
   If you have used an AI tool to help refine your work, state which tool and what prompts were given. You can also provide a URL of the chat log instead (please make sure the URL is accessible).

   If you have not used an AI tool to help refine your work, please explicitly state so.

# Example Research Question

This is an example to demonstrate what a research question is and what we mean by different "ways" to explore the question. Let's assume that our theme of the semester was "Education at UBC".

You can take two approaches to brainstorming research questions

1. Ask yourself what you wonder about UBC. Is there anything that you ask yourself when you have to determine what kinds of courses you want to take? Do you have any questions about the way pedagogical decisions are made at UBC?

2. Look into what datasets UBC offers for public access and then scan through the data to see if there are any strange patterns or outliers that stand out. This is often referred to as "exploratory data analysis".

Either way, at some point you will come up with a question. You will often find that questions brainstormed at this stage still need refining.

For example, imagine this was the first version of the research question "Did the pandemic have an impact on UBC students?".

This is a good starting point but the question still needs work. Here would be some things that would need further refinement:

1. What does "impact" mean? Are you referring to the psychological, financial, emotional, or educational side of things?

2. Regardless of what you choose for "impact", how would you measure it and **do you have the data that would allow you to answer the question**?
    a. For example, assume we chose "educational impact". Now the question would be how are you defining educational impact? By the number of credits attempted each semester? Overall GPA? The number of misconduct cases handled by the university? All of these are valid ways to explore this question. Each team member may take one of these directions to use for their project contribution.

3. What are you comparing against? UBC students from prior to the pandemic? High school students? SFU students? Again, this is a question of "do I have the data?".

4. UBC has many students. It would be wise to narrow down the search space to make it easier for you to sift through the data. You can narrow down your study population by isolating it by year level (e.g., first years, second years, etc.) and/or by major and/or by faculty/school.

After refinement, you may come up with a research statement that looks like this:

> We will compare the educational impact of the pandemic on domestic versus international first year Arts students who enrolled at UBC in 2021 (pandemic learn-from-home year) and 2023 (in-person learning).
>
> Educational impact will be explored in three ways:
>
> 1. The average number of attempted credits for each school year (defined as the courses taken between September and April).
> 2. The number of individuals who obtained a GPA greater than the average in their cohort for each school year
> 3. A survey sent to UBC faculty to solicit feedback on how teaching has changed pre, during, and post pandemic. These responses will be coded for common themes.

For this class, you will **not** be asked to run a survey or to gather data to create your own dataset. We want you to use **existing** datasets from **reputable** sources.

The third research question is given as an example that you can have research questions that deal with qualitative data instead of purely quantitative questions. Depending on your datasets, you may have qualitative data. You must have **at least one** quantitative question involving the summation/averaging/division/subtraction/multiplication of data.

You can have **at most one** qualitative question. Qualitative data analysis often takes much more time than quantitative analysis so if you are choosing a qualitative question, please do consider your own commitments and schedule first.