# DSCI 542 Lab 1

**Communicating in Data Science**

## Lab Assignment Instructions

In this course, we are using a labor-based grading approach. This means that each task in the lab is graded based on completion rather than evaluation of quality. If a question is completed according to the specifications provided such as meeting the word limit, following the required format, or staying within the allotted time you will receive full credit for that question.

When submitting your lab, please be sure to:

- Include link to your GitHub repository containing your work.
- Submit a the `.qmd` file along with a rendered version (either `.html` or `.pdf`) of your lab to Gradescope.
- On Canvas, submit the link to your GitHub repository containing your work for peer review.

**GitHub Repository:** https://github.com/PAT0216/DSCI542_labs

## Question 1: Asking the right questions (12 pts)

rubric = {completion: 12}

An essential part of communication in data science is asking the 'right' questions. A good question should stem from a real business problem and reflect the direction of the analysis. As you will see later on in capstone projects, you will be often put in a situation where you need to translate a client's business interests into a concrete & specific question that determine the direction of the whole project. For example, the client might want to **describe** some existing patterns in the data, or they want to **make inference** to a wider population, or they want to **predict** some outcomes, or they might want to understand if certain interventions have a **causal impact** on some outcomes. You will need to read between the lines and identify what the client's interests are to formulate a meaningful question.

In this exercise, you will be formulating different types of questions (**descriptive, inferential, prediction, causal**) to given problems.

| Type | Description | Question |
|---|---|---|
| Exploratory | A question that asks if there are patterns, trends, or relationships within a single data set. Often used to propose hypotheses for future study. | Does political party voting change with indicators of wealth in a set of data collected on 2,000 people living in Canada? |
| Predictive | A question that asks about predicting measurements or labels for individuals (people or things). The focus is on what things predict some outcome, but not what causes the outcome. | What political party will someone vote for in the next Canadian election? |
| Inferential | A question that looks for patterns, trends, or relationships in a single data set and also asks for quantification of how applicable these findings are to the wider population. | Does political party voting change with indicators of wealth for all people living in Canada? |
| Causal | A question that asks about whether changing one factor will lead to a change in another factor, on average, in the wider population. | Does wealth lead to voting for a certain political party in Canadian elections? |

Source: https://datasciencebook.ca/intro.html#asking-a-question

For questions 1.1-1.3, FORMAT YOUR ANSWER LIKE THIS:

| Type | Question |
|---|---|
| Exploratory | |
| Predictive | |
| Inferential | |
| Causal | |

### 1.1 Problem 1: Real estate marketing

A digital real estate marketing firm is hiring you as a data science consultant. It offers a comprehensive list of for-sale properties, as well as the information and tools to make informed real estate decisions.

Return users are consumers who visit their website more than once. They may return for a variety of reasons such as to discover new homes, search for open houses, know about price reductions, check changes in status of previous homes of interest. These consumers form a very important segment of users since these are the ones we can expect to take more interest in our brand, be more engaged, loyal and, finally convert to a lead.

In this project we seek to build a data science/machine learning framework to find users that are likely to return and to develop strategies to encourage them to return to our website.

**ANSWER**:

| Type | Question |
|---|---|
| Exploratory | Are there any patterns in browsing behavior (like time spent or pages viewed) that differ between users who return and those who don't? |
| Predictive | Based on a user's first visit behavior, can we predict whether they'll come back to our website within the next 30 days? |
| Inferential | Do users who engage with price reduction alerts generally show higher return rates compared to the broader population of our website visitors? |
| Causal | Does sending personalized "homes you might like" emails actually cause users to return to our website more often? |

### 1.2 Problem 2: Sport analytics

A professional basketball team is hiring you as a data science consultant. A primary mandate of this team is to inform decisions across several areas of Basketball Operations, including decisions about athletes' physical preparation, injury prevention, return to sport, player identification/scouting, performance analysis, and opposition scouting.

The team collects a large amount of optical tracking data that records the x/y coordinates of all players and the ball at a frequency of 25 frames per second, while the event data records "on-ball" actions throughout the course of a match such as passes, shots.

In this project, we seek to apply a data science framework to enhance our players performance and develop effective squad formations that increase our chance of winning games.

**ANSWER**:

| Type | Question |
|---|---|
| Exploratory | What movement patterns or passing sequences tend to happen more often in games that we win versus games we lose? |

| Type | Question |
|---|---|
| Predictive | Given the current lineup and play positioning, what's the expected probability of scoring on this possession? |
| Inferential | Do players who follow our new training regimen show improved shot accuracy across all players in professional basketball? |
| Causal | Does implementing a zone defense strategy actually lead to fewer opponent points compared to man-to-man defense? |

### 1.3 Problem 3: Learning analytics

Coursera is hiring you as a data science consultant. The company collects a large amount of event log data from learning management systems, in addition to academic records and personal information. Suppose that while these massive open online courses (MOOCs) have large enrollment, the completion rate has been observed to be rather small.

In this project, we are interested in identifying students who are at-risk of dropping out their course and how can we better retain these students.

**ANSWER**:

| Type | Question |
|---|---|
| Exploratory | Are there common patterns in how at-risk students interact with course materials compared to those who complete the course? |
| Predictive | Can we predict which students are likely to drop out in the next two weeks based on their current engagement metrics? |
| Inferential | Is the relationship between video watch time and course completion rate generalizable to all MOOC learners, not just those in our sample? |

| Type | Question |
|---|---|
| Causal | Do reminder notifications actually cause students to be more likely to complete their courses, or do engaged students just happen to respond to them? |

## Question 2: Outlining (6 points)

rubric = {completion: 6}

In this course, we will spend a lot of time reflecting on how to clearly communicate ideas in data science. To begin, we would like you to choose any data science topic that interests you and create an outline for a short presentation on that topic.

At this stage, do not worry about making the outline perfect. For now, the goal is simply to sketch out the main sections and points you would include if you were giving a presentation about this topic.

> *Note: Please think carefully about the topic you choose. This choice matters because it will not only shape the outline you create now, but will also be woven into different parts of the course as we practice communication skills. In fact, there's a strong chance that the topic you select here will become the foundation for the presentation you deliver later.*
>
> *Note: Used GenAI Gemini 3 pro for proofreading and formatting of content.*

Topic: Paper Trader AI - Building an Automated Trading System

I. Introduction

- What is paper trading?
- Why it's useful for learning

II. The Problem

- Emotional trading decisions
- Time-consuming to monitor markets
- Risky to test strategies with real money

III. My Solution

- Automated trading bot
- Three different strategies (Momentum, ML, LSTM)
- Daily automated execution

IV. Demo

- Live dashboard walkthrough V. Results
- Performance comparison
- Key learnings

VI. Conclusion

- Main takeaway
- Limitations

### Question 3: Explain complex concepts to non-technical audience (10 points)

rubic = {completion: 20}

Explaining complex data science concepts to non-technical audience is a valuable skills set for any data scientists. By doing so, you can clearly communicate the results of their analysis and the insights they have gained, helping non-technical audiences to **understand the significance and implications of their work**. By writing clearly and concisely, data scientists can ensure that their ideas and insights are understood by their colleagues and that they can **work together effectively**. Finally, it can help you **build credibility and trust with your colleagues**, and demonstrate the value and relevance of your work.

Here are some tips for effective writing data science communication:

- **Be clear and concise**: Avoid using jargon or overly technical language, and aim to explain concepts in a simple and straightforward manner.
- **Use examples**: Illustrate your points with examples or case studies to help make the material more relatable and easier to understand.
- **Break down complex concepts** If you are explaining a complex concept, break it down into smaller, more manageable pieces and explain each piece separately.
- **Use visual aids**: Use charts, graphs, or other visual aids to help clarify and illustrate your points.
- **Ask for feedback**: Review your writing carefully to ensure that it is accurate and well-organized, and ask for feedback from others if needed.

In this exercise, you are tasked to explain your topic of choice to a non-technical audience (a type of activity that you will find yourself doing frequently later on when you work as a data scientist). For this exercise, you should be aiming for a character count of around 2500 (+ or - 500) characters, not including code, captions, references, etc. Please consider the amount of time it will take your audience to read your post (about 5-10 minutes). Consider the following points before diving into your write-up:

- Identify and highlight technical jargon and complex terms that might not be easy to understand by non-technical audiences. Please keep in mind the audience that you are writing for.
- Refine the outline for your write-up.
- Rewrite and explain the concept in a clear and concise way, creating additional visual aids if necessary.

  *Note: Used Gemini 3 Pro for proofreading and formatting of content/smoothing upto a certain extent.*

## Paper Trading and Algorithmic Trading: The Why and How.

What if you could Practice Investing Without Losing Money?

Most people are curious about investing but intimidated by the risk of losing money. This is where paper trading comes in. It is a method of simulating trades using fake money while tracking real market prices. Think of it as training wheels for the stock market. You get all the learning experience without any of the financial pain.

However, one major issue with traditional trading is that humans are emotional. When the market drops, people tend to panic and sell. When a stock is skyrocketing, greed kicks in and they buy at the peak. These emotional decisions often lead to poor results. Studies show that most individual investors actually underperform the market because of these behavioral mistakes.

This is why algorithmic trading has become so popular. Instead of making decisions based on gut feelings, an algorithm follows a set of predefined rules based purely on data. The algorithm doesn't panic. It doesn't get greedy. It simply executes the strategy it was programmed to follow.

One approach is called momentum trading, which is surprisingly simple: stocks that have been going up tend to keep going up, and stocks that have been falling tend to keep falling. However, a more advanced approach involves using Machine Learning to find hidden patterns in the data.

One popular machine learning method is called XGBoost. To understand how it works, imagine you are trying to guess whether it will rain tomorrow. You might ask a series of yes/no questions: Is it cloudy? Is the humidity high? Was it raining yesterday? Each question helps narrow down the answer. XGBoost works in a similar way. It builds hundreds of these "question trees" and combines their answers to make a final prediction. Each tree learns from the mistakes of the previous one, gradually improving accuracy. In trading, XGBoost can analyze factors like recent price movements, trading volume, and market trends to predict whether a stock will go up or down.

The beauty of machine learning is that it can spot patterns that humans might miss. However, it is not perfect. Just because a model worked on past data doesn't mean it will work in the future. Markets change, and what worked yesterday may fail tomorrow. This is why testing strategies using paper trading is so important. You can see how the model performs in real conditions without risking actual money.

Additionally, it is important to understand transaction costs. Every time you buy or sell a stock, you pay a small fee. These fees seem tiny, maybe a few dollars, but they add up fast when you're trading frequently. Many seemingly "profitable" strategies actually lose money in practice because the fees eat away at the returns.

To summarize, paper trading allows anyone to experiment with investing ideas safely. Machine learning, like XGBoost, can help find hidden patterns and make predictions. And understanding transaction costs reveals whether a strategy is truly profitable or just an illusion.

## Peer Feedback (to be completed next week)

After the submission deadline, you will review your group member's write ups.

Peer review allows you to receive feedback on your writing from multiple peers, which help you see your work more objectively and identify areas where they can revise and improve their writing. It also fosters a sense of community and a safe environment to collaborate and help each other with constructive feedback.

Read the write-ups from Question 3 carefully, paying attention to the content, organization, and style of the writing. Consider the following questions as you review:

- Is the content of the write-up accurate and well-researched?
- Is the structure of the write-up clear and logical?
- Are the ideas in the write-up effectively communicated and easy to understand?
- Is the writing style of the write-up engaging and appropriate for the audience?

Write a review for each of the write-ups, addressing the above questions and any other issues you feel are relevant. Use specific examples to support your points. Reviews should be **at least 250 words** in length. Please be sure your feedback is positive, constructive, and focused on helping your peers strengthen their work.