DATA 603 Project Checkpoint

Due: Tuesday November 19th by 9pm

Objective

The point of this first checkpoint is to help you get thinking about possible data you would like to use for the project as well as a possible research topic that you would like to try to answer with this data.

While I don't expect you to have every aspect of your data and/or research question in place by this point, any info you give me will help me determine if the data you plan on using (and the question you'd like to try to answer) are appropriate for the final project. This checkpoint is basically a way for you to "double check" your idea with me to make sure it will actually work with the course material!

What You Need to Do for Project Checkpoint

Please answer the following questions and submit your answers to Dropbox via D2L by 9pm of the class on Tuesday Nov 19th .

You are working in a group of **3-4 people per group**, please—only one set of questions needs to be completed per group (just make sure all group members' names are on the sheet!).

1. Name of all group members.
2. If you have a research topic in mind for this project, please briefly describe the topic and any background info I may need to understand the topic.
3. Please briefly describe the data you have (or plan to acquire) to help answer the research topic above. Include: what type of variable or variables are included (quantitative, qualitative, etc.), how the variable or variables are measured (the measurement scale), and any other general info you may have on the variable(s).
4. Is this your own data set (or the data of someone in the group) or is it "open" or "shared" data?

Project Checkpoint template is provided in D2L under Project information folder

The following pages contain some additional information that might be useful as you start the process of thinking about the project.

Additional Information

Below, **I have included a brief summary of Statistical Modelling which we will be covering in this course**. This is to help you get an idea of what types of modelling we will be learning so that you can get an idea of what type of analysis might be appropriate for the data you're planning on using. I've also included an example research topic for each modelling that would give you an idea of the type of research topic that you would like to analyze.

Statistical Modelling Covered in DATA 603

1. **Multiple Linear Regression**: a technique to model the relationship between two or more explanatory variables (independent variables) and a response variable (dependent variable) by fitting a linear equation to observed data. It is used when we want to predict the value of a variable based on the value of two or more other variables.

- **Example research topics :**

  - Analysis of residential real estate prices depend, in part, on property sizes and number of bedrooms.

    Response variable: *Residential real estate price (quantitative variable).*
    Explanatory variables: *Property size (quantitative variable) and Number of bedrooms (quantitative variable).*

  - Relationship between television advertising expenditures and print media advertising expenditures on sales revenue.

    Response variable: *Sales revenue (quantitative variable).*
    Explanatory variables: *Television advertising expenditures (quantitative variable) and Print media advertising expenditures (quantitative variable).*

  - Examine the price of diamonds based on their characteristics.

    Response variable: *Price of diamonds (quantitative variable).*
    Explanatory variables: *Carat (quantitative variable), Cut (qualitative variable), and Clarity (qualitative variable).*

- Forecast health insurance charges based on several key characteristics.

  Response variable: *Charges (quantitative variable).*
  Explanatory variables: *Age (quantitative variable), Sex (qualitative variable), and Clarity (qualitative variable), Body Mass Index (BMI) (quantitative variable), Children (quantitative variable), Smoker (qualitative variable), Region (qualitative variable).*

2. **Logistic Regression:** a technique to model the relationship between two or more explanatory variables (independent variables) and an outcome (dependent variable) which an outcome is measured with a dichotomous variable (two possible values when observed: yes vs. no, positive vs. negative, died vs. alive, etc)

- **Example research topics :**

  - Predict if a given credit card transaction is fraud or not.

    Response variable*: Credit card transaction (dichotomous variable).*
    Explanatory variables: *time (quantitative variable), amount (quantitative variable), and length (quantitative variable).*

  - Model and predict if a given specimen is benign or malignant.

    Response variable*: Mass of tissue (dichotomous variable).*
    Explanatory variables: *thickness (quantitative variable), Cell Size (quantitative variable), and Cell Shape (quantitative variable).*

  - Compute a prediction probability score of a customer who may default on a loan.

    Response variable*:* Credibility *(dichotomous variable).*
    Explanatory variables: *Work Experience (quantitative variable), Gender (qualitative variable), Level of Education (qualitative variable), and Salary(quantitative variable).*

3. **Design of experiment:** a technique to determine the relationship between factors affecting a process and the output of that process.

- **Example research topics :**

  - Analyze the average annual per capita beer consumption across 3 regions of the world: Asia, Europe, and America.

    Response variable*: Amount of beer (liters).*
    Explanatory variable(s): *regions (*Asia, Europe, and America*)*

  - Investigate the relationship between a consumer's transaction history (levels: long and short) and an employee's statement of thanks (levels: yes and no) on a consumer's repurchase intent which was measured using a 9-point scale.

    Response variable*: consumer's repurchase intent (9- point scale).*
    Explanatory variable(s): *consumer's transaction history and employee's statement of thanks.*

Open Data Sources: Below are some links to access open data sets that are freely available to everyone to use and republish.

-University of Calgary Data Sources for Data Science	https://library.ucalgary.ca/datasources
-Open Calgary	https://data.calgary.ca/
-Government of Alberta's Open Data Program	https://www.alberta.ca/open-government-program.aspx
-Government of Canada's Open Data Portal	https://open.canada.ca/data/en/dataset?organization=statcan
-Kaggle's Open Data Portal	https://www.kaggle.com/datasets
-Google Dataset Search	https://toolbox.google.com/datasetsearch
-GitHub's Open Data	https://github.com/collections/open-data