**Using Data Science to Understand the North American Data Science Labour Market:   
A Canada-U.S. Comparison**

**Business Understanding:**

The initial phase is to understand the project's objective from the business or application perspective. Then, you need to translate this knowledge into a machine learning problem with a preliminary plan to achieve the objectives.

[P-O]: Kaggle has been surveying professionals working in the field of data science for a few years to improve our understanding of the dynamics of the job market. The datasets available are covering dozens of countries and territories across the world. While the Kaggle community has offered many insights around the case of the United States, there remains a lot to be discovered about data science in Canada. By shedding light specifically on the Canadian data science job market and comparing it with its counterpart in the United States, this data science project will help employers, employees, and candidates better understand the skills needed to become a data scientist, the job titles they are likely to have, as well as whether these skills and titles vary across the two countries.

We will first provide descriptive statistics and visualizations for both countries to gain insights into the current state of the data science labour market in Canada and the United States. Then, we will build models that will predict whether a respondent is a Canadian or American data scientist based on the main features of each of the two subsamples. More specifically, we will build a decision tree classifier, a support vector machine, a logistic regression model, and a KNN model. We will then evaluate models and compare them using metrics including F1 scores, the Jaccard index, and log loss as appropriate.

This project is especially relevant in a time where work-from-home arrangements are becoming prevalent. The current pandemic has brought about major restructurations in the industry and companies are now looking to hire data scientists across the globe. Classifying the distinct features of a specific labour markets should be a valuable enterprise.

**Data understanding:**

In this phase, you need to collect or extract the dataset from various sources such as csv file or SQL database. Then, you need to determine the attributes (columns) that you will use to train your machine learning model. Also, you will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.

[P-O]: As mentioned above, the dataset that we will use comes from Kaggle. The company conducts an annual survey and this one has been conducted in 2019, from October 8th to October 28th. Respondents were “found primarily through Kaggle channels, like [their] email list, discussion forums and social media channels” (see: <https://www.kaggle.com/c/kaggle-survey-2019/data?select=multiple_choice_responses.csv>). In total, there were 19 717 respondents from 171 countries and territories across the globe.

For the features of the models, we need to select those that are most appropriate to predict the outcome of interest. Here, we are interested in knowing what distinguishes data scientists from Canada and from the United States. So, we select as the outcome of interest the variable “Q3”, which corresponds to the country of the respondent.

Next, we select a few sociodemographic variables likely to be important in distinguishing the two groups. For instance, it is possible that the age structure of data scientists in the U.S. is different than Canada’s. Otherwise, perhaps there are proportionally more women working in data science in Canada than in the U.S. As for education, perhaps there are differences regarding the level attained among data scientists in both countries.

Additional features of interest will be included. After reviewing the entire set of questions of the survey, we hypothesize that there may be differences between the two cases studied when it comes to the following features:

* Age (“Q1”)
* Gender (“Q2”)
* Reminder: the variable “country”, Q3, is the outcome of interest so we include it along with the rest of our predictors into our new data frame
* Level of education (“Q4”)
* Job title (“Q5”)
* Company size (“Q6”)
* Data science team size (“Q7”)
* Incorporation of machine learning methods (binary variable) (“Q8”)
* Salary (annual) (“Q10”)
* Primary tool used (“Q14”)
* Prior experience coding (in years) (“Q15”)
* Programming languages used on a regular basis (“Q18”… This one is a little tricky because it was a multiple answers question. The question has been broken down into twelve distinct variables, ranging from “Q18\_Part\_1” to “Q18\_Part\_12”)
* Number of years using machine learning methods (“Q23”)

**Data Preparation:**

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

[P-O]: As done by Kaggler Shivam Bansal (see <https://www.kaggle.com/shivamb/spending-for-ms-in-data-science-worth-it>) in their analysis of whether it is worth pursuing a formal degree in data science to work as a data scientist, we exclude respondents who declared being students or unemployed at the time they filled out the survey questionnaire. We are going to focus on the rest of the sample, which comprises all employed respondents wearing different data science job hats.

In terms of pre-processing, we take the following steps:

* Remove the first row since it contains question labels
* Check the number of missing values for each column and use this information to make a decision as to which variables are or are not usable. We decide not to use questions after Q23 since the proportion of missing values is too high.
* Create a dataframe with only the features selected (as described above in the “Data understanding” section) and only [Canadian | American] respondents.
* Reformat variables so they can all be used for modelling.
  + Transform Q1 (age) into an ordinal variable ranging from 0 to 10 since there are 11 age categories ranging from 18-21 to 70+
  + Transform Q2 (gender) into a binary variable where “Male” = 0, “Female” = 1, “Prefer not to say” = 2, and “Prefer to self-describe” = 2
  + Transform Q3 (country) into a binary variable where “United States of America” = 0 and “Canada” = 1

**Modeling:**

In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. You can select SVM, XGBoost, decision tree, or any other techniques. You can select a single or multiple machine learning models for the same data mining problem. At this phase, stepping back to the data preparation phase is often required.

**Evaluation**:

Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.

**Deployment:**

The deployment phase requirements vary from project to project. It can be as simple as creating a report, developing interactive visualization, or making the machine learning model available in the production environment. In this environment, the customers or end-users can utilize the model in different ways such as API, website, or so on.