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

Introduction

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 models. We will then evaluate models and compare their accuracy.

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 specific labour markets should be of interest to a range of actors including the following: 1) data science candidates willing to understand how they compare with current data scientists in their country as well as their neighbouring country, 2) recruiters who need to understand the features of the data science labour market in North America, and 3) organisations and companies that want to evaluate their current level of data science capital and maturity from a comparative perspective.

The Data

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 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. The subsamples comprise 2489 American respondents and 355 Canadian respondents.

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 several differences between the two cases studied when it comes to the following features:

- Age ("Q1")
- Gender ("Q2")
- (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, which we will separate later)
- Level of education ("Q4")
- Job title ("Q5")
- Company size ("Q6")
- Data science team size ("Q7")
- Incorporation of machine learning methods ("Q8")
- Salary (annual) ("Q10")
- Primary tool used ("Q14")
- Number of years using machine learning methods ("Q23")

Methodology

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 hats.

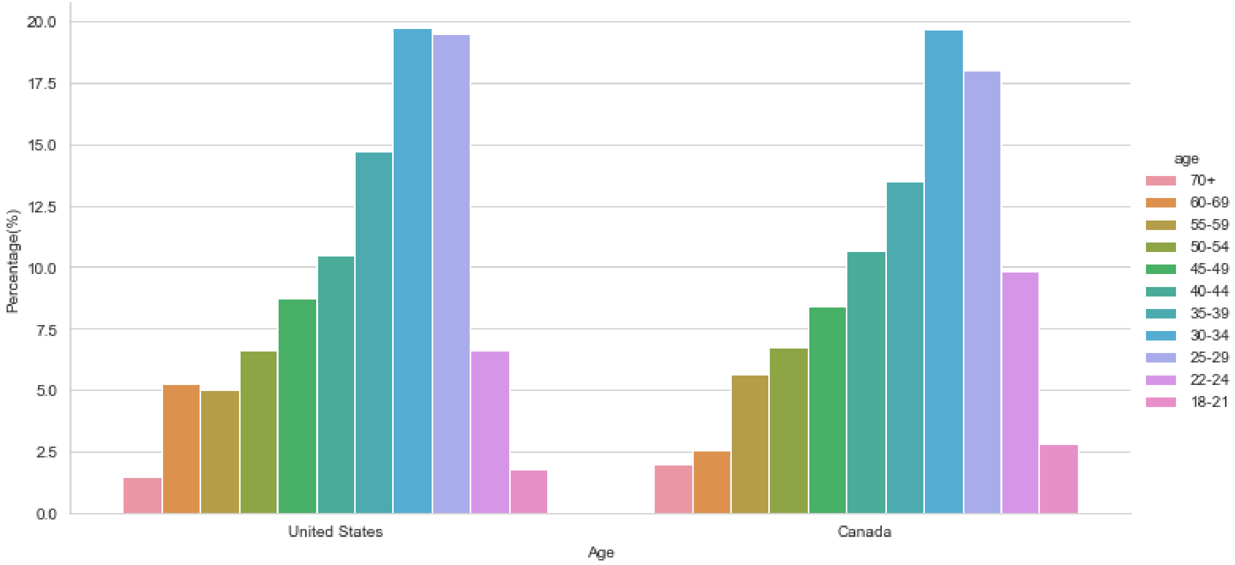
In terms of pre-processing, we take the following steps:

- examine the full dataset
- look at the question labels of each questions to ensure a thorough understanding of the dataset
- select features of interest
- drop the first row of the dataset, which contains question labels
- examine the data types to ensure they are properly formatted
- exclude students and unemployed individuals from the dataframe since we are seeking to examine the structure of the labour market as it stands, not as it may become once some respondents have landed a position in the field
- create a new dataframe with a reduced number of features that are likely to be important in distinguishing between the two countries studied
- select a subsample to reduce the sample so it includes only Canadian and American respondents to fit with the objectives of our analysis
- relabel the columns so we can intuitively understand the dataframe when looking at it in Python
- deal with missing values for modelling: for this project, we are simply going to drop the rows containing missing values since we have a dataset of sufficient size for our purpose.
- Transform the categorical variables of age, ML_yrs, and salary into ordinal variables. It makes sense to do so, but they are currently treated as nominal variables (*i.e.* without any ordering).

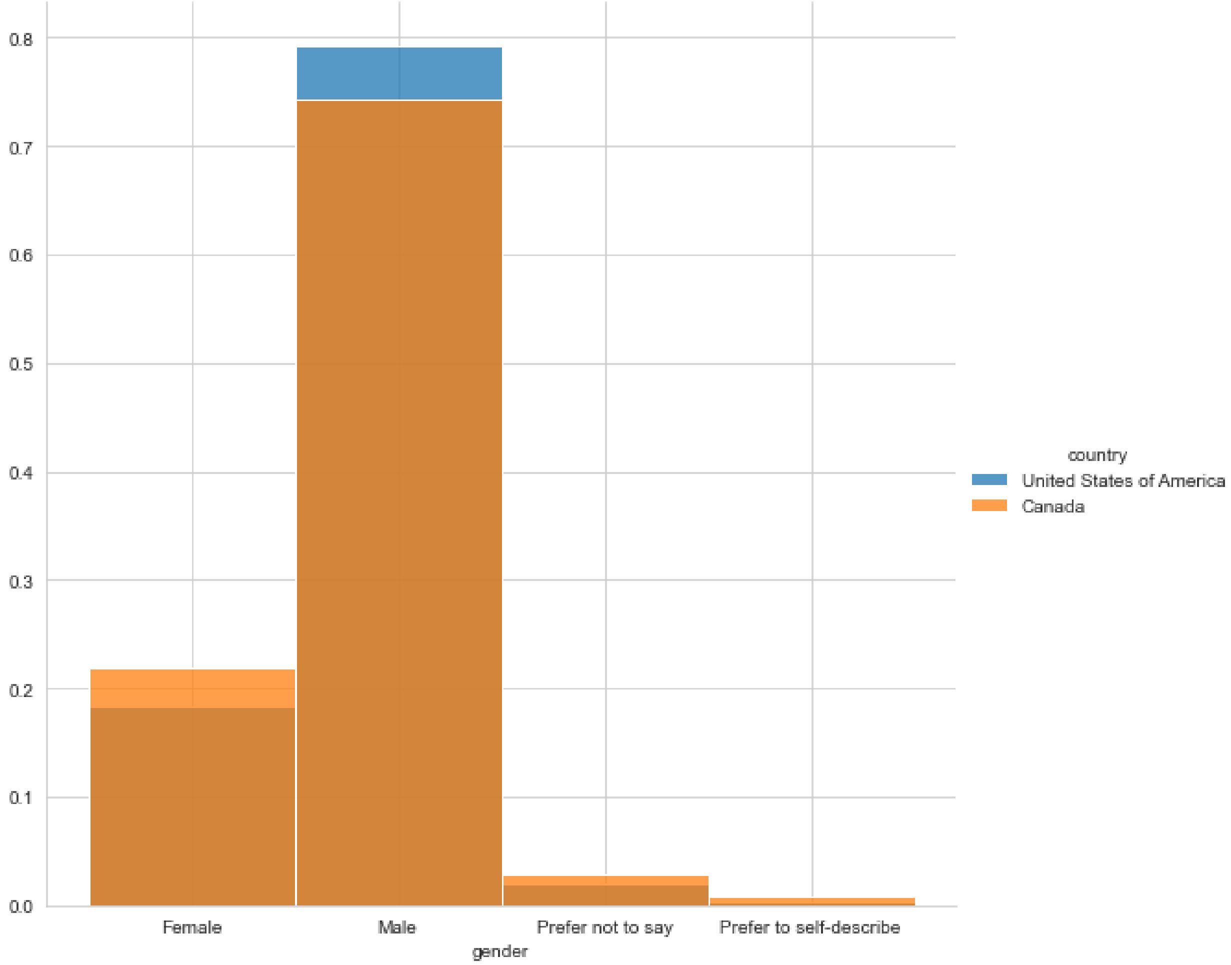
Results

Exploratory Data Analysis and Visualization

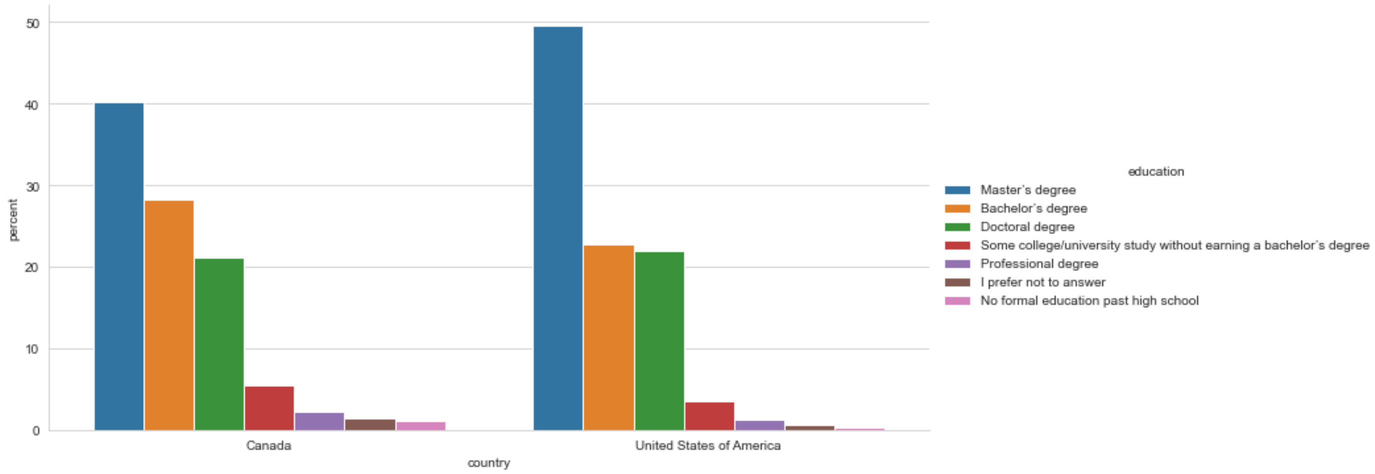
Starting with the first of our ten variables selected as features for our models to come, we observe the age distribution of data scientists in Canada and the United States. As we see, both countries have a very similar distribution, with two noticeable differences: the U.S. has about 2.5% more data scientists in the 60-69 age range, and Canada has about 4% more data scientists in the 22-24 age range. The two most common age categories in both countries are the 25-29 years old and the 30-34 years old. Both countries have an average age range of 30-34 years old.



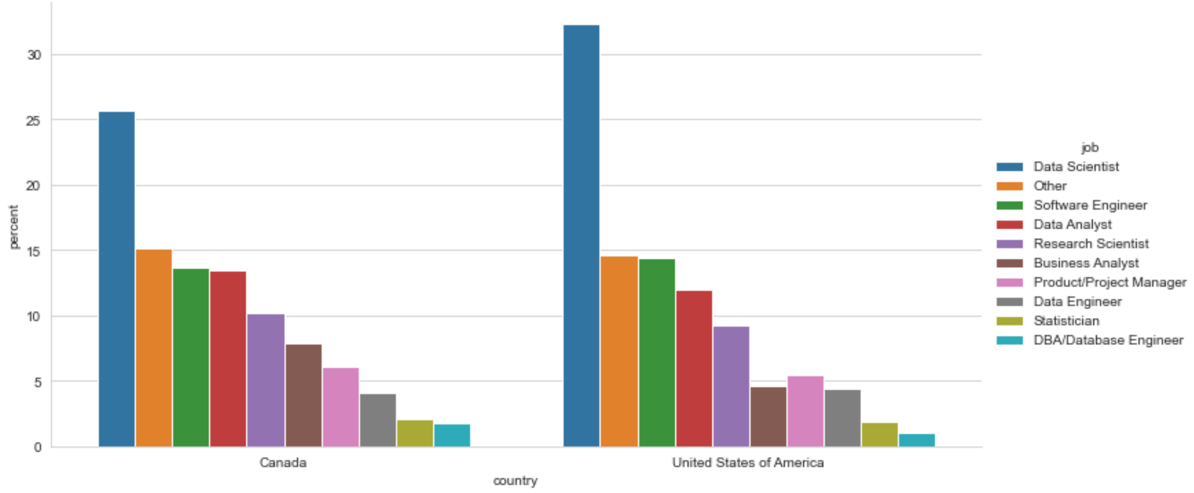
Regarding gender, there really is not much to say except that the data confirm what is already widely known: men form the overwhelming majority of the data science professions, both in the U.S. and Canada.



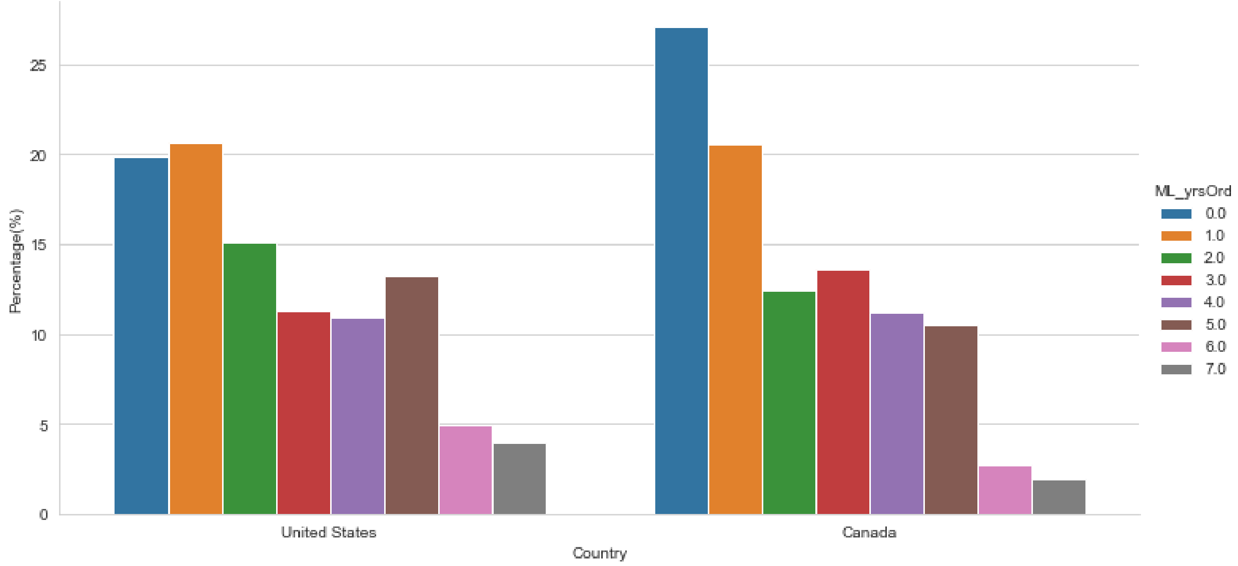
More differences begin to appear in the next features. In terms of education, for example, the United States clearly have a higher proportion of professionals holding a Master's degree in comparison with Canada (about 10% more). Canada, however, has more data scientists holding a Bachelor's degree (about 5% more).

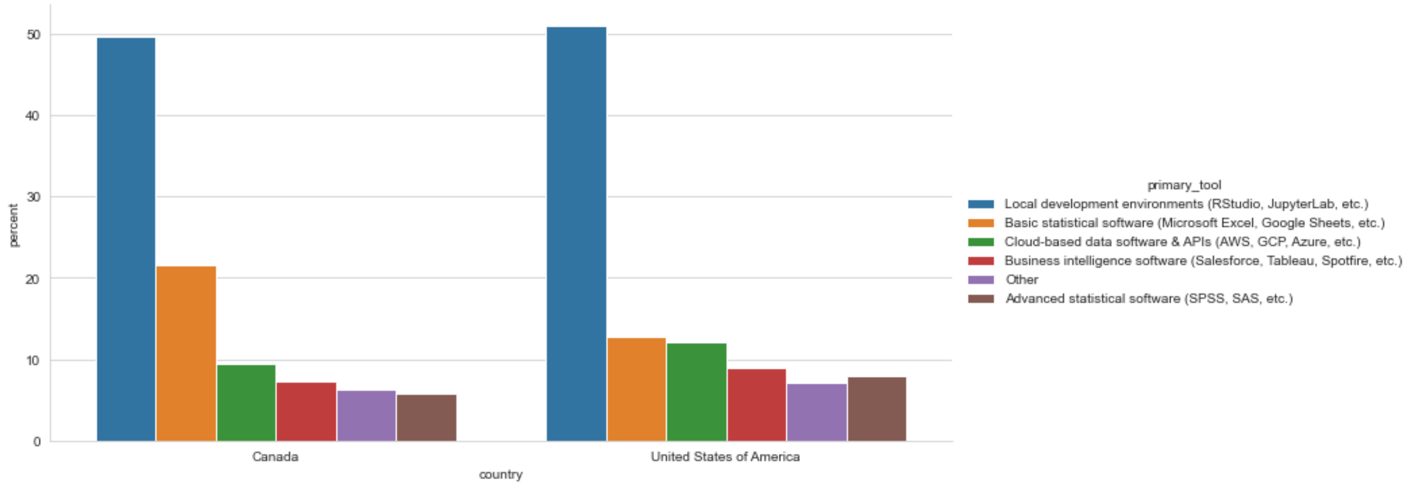


Likely a consequence of the above-mentioned difference, the U.S. has more professionals wearing the "data scientist" title than Canada. Most of the distribution is similar when it comes to other titles, with the noticeable exception that Canada has about 3% more "business analysts" than the U.S. We will see further below in the heatmap of associations between nominal variables, however, that education does *not* have a strong statistical association with job title.

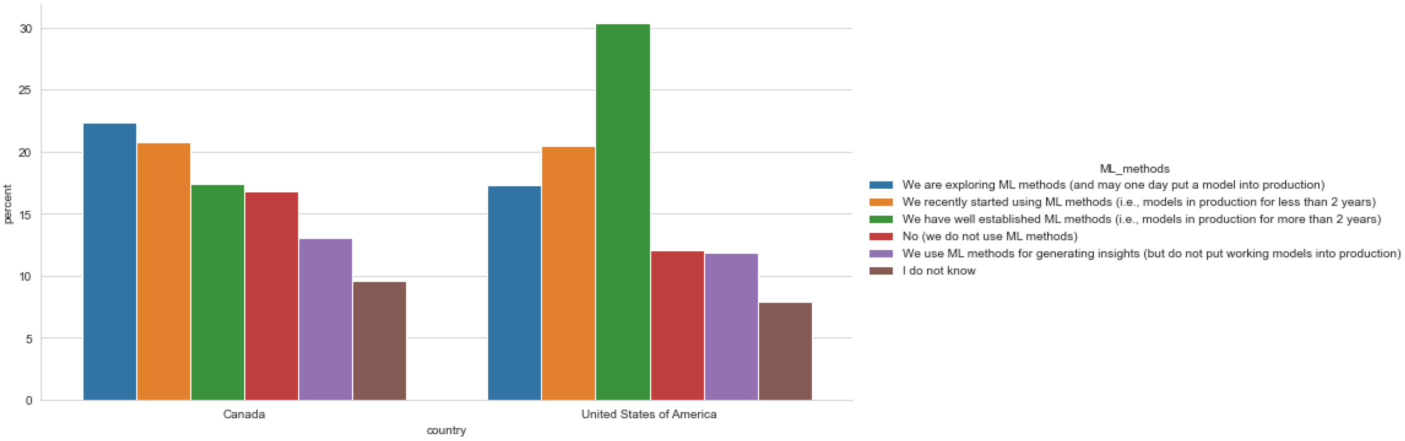


Next, we see in the figure below that Canada has more junior data scientists than the United States. Measured in number of years doing machine learning, where 0= “<1 year”, 1=”1-2 years”, 2=”2-3 years”, 3=”3-4 years”, 4=”4-5 years”, 5=”5-10 years”, 6=”10-15 years”, and 7=”20+years”.

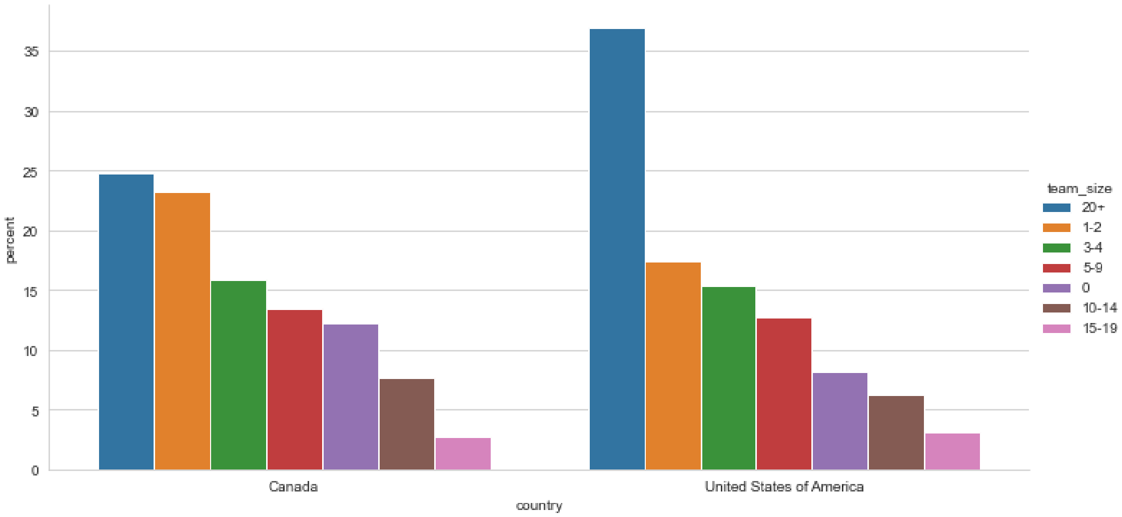




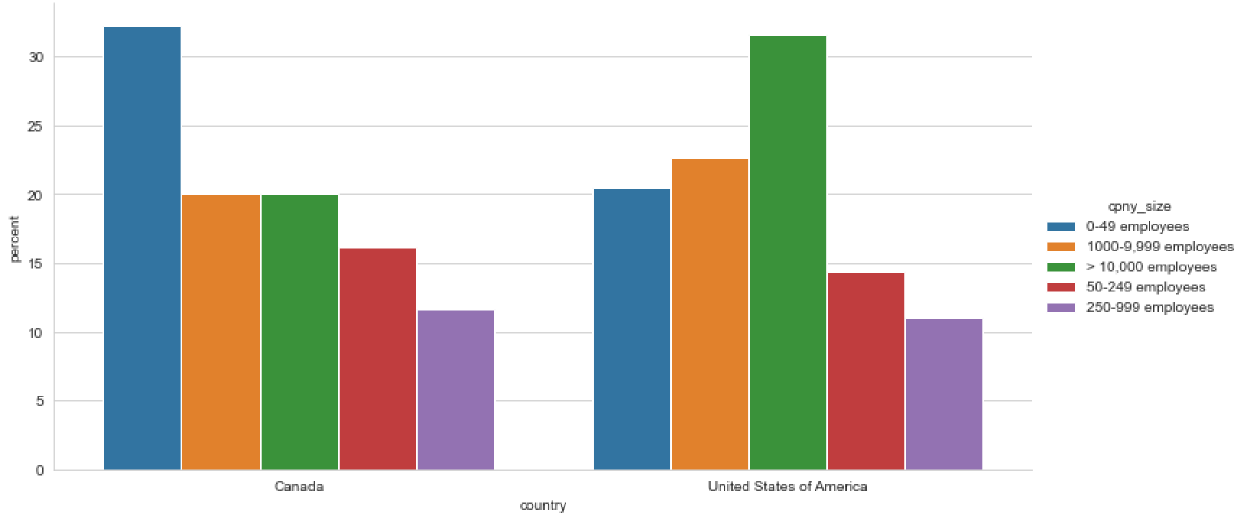
Above, we observe that a majority (over 50%) of American data scientists are using local development environments are their data science primary tool whereas it is just short of a majority who do so in Canada. There is also a noticeable difference, about 10% more people using basic statistical software such as Microsoft Excel in Google Sheets in Canada than in the United States. This points toward the notion that the Canadian data science market might a little less mature than the American one. Evidence confirming the latter hypothesis is found in the next figure, which shows that while over 30% of the American data scientists surveyed have "well established ML methods (*i.e.* models in production for more than 2 years)", it is the case for only about 17% of Canadian data scientists. As a consequence, there are slightly more Canadian data scientists who have only recently started using ML methods or do not use ML methods at all.



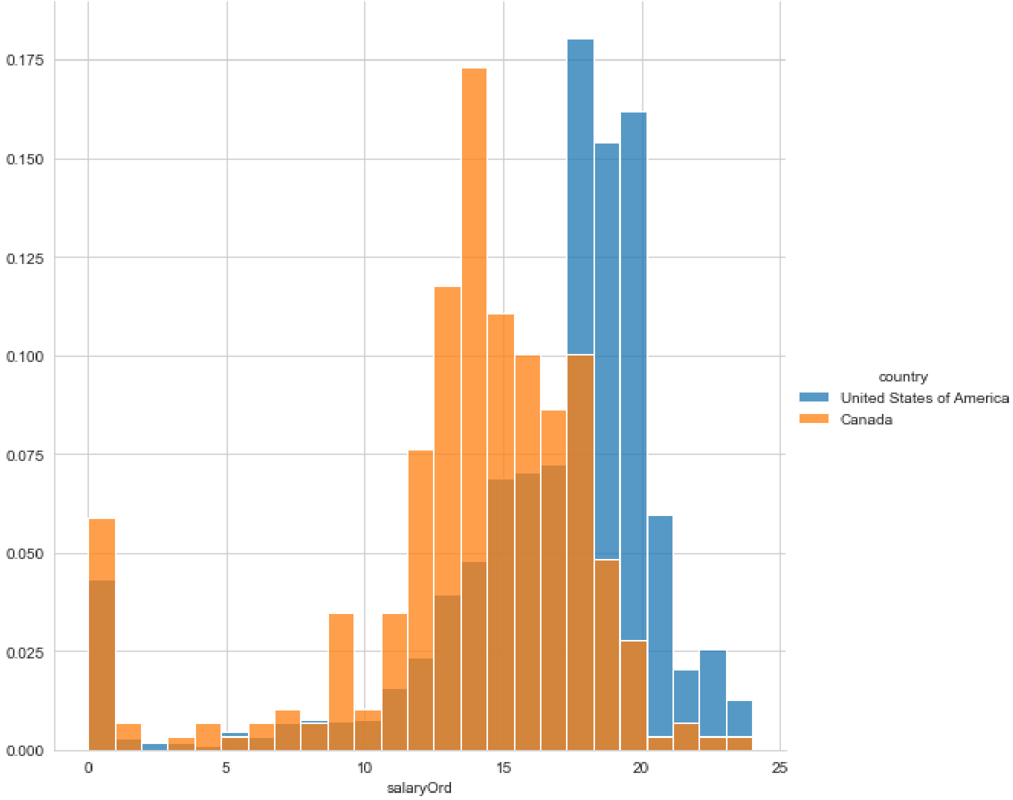
Unsurprisingly, the next figure below shows that American data scientists are working in teams of greater size than their Canadian colleagues. Team sizes of more than 20 are indeed about 17% more frequent in the U.S. than in Canada.

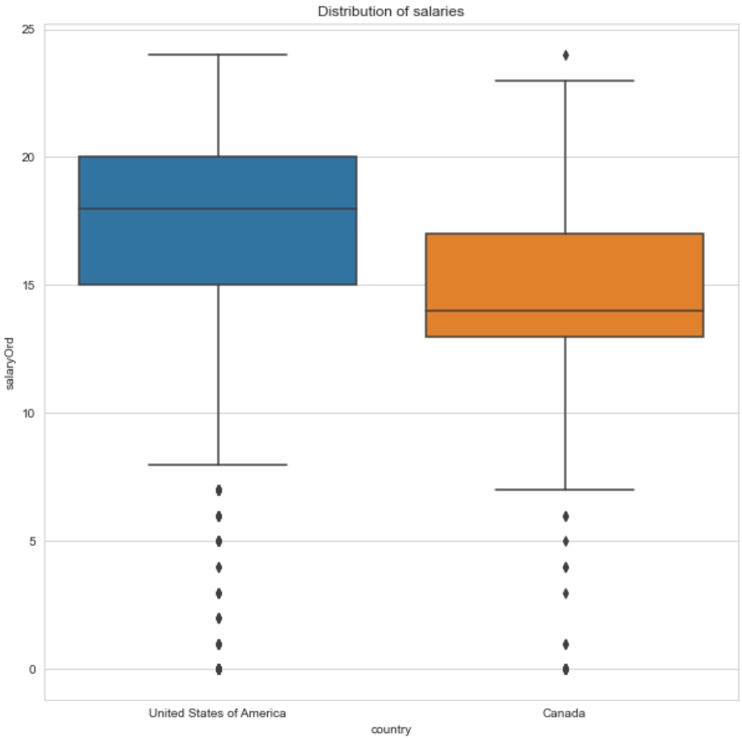


Team sizes as a feature distinguishing the U.S. data science market from its Canadian counterpart seems to go hand in hand with company size. Just like the U.S. data scientists are on average working in larger team sizes than Canada, the former are working in companies of larger size as well. The largest share of American respondents (~32%) American respondents indicated they were working in a company of over 10,000 employees while the largest share of Canadian respondents (~33%) indicated they were working in a company of 0-49 employees. This might be one of the major differences observed so far between the two markets.

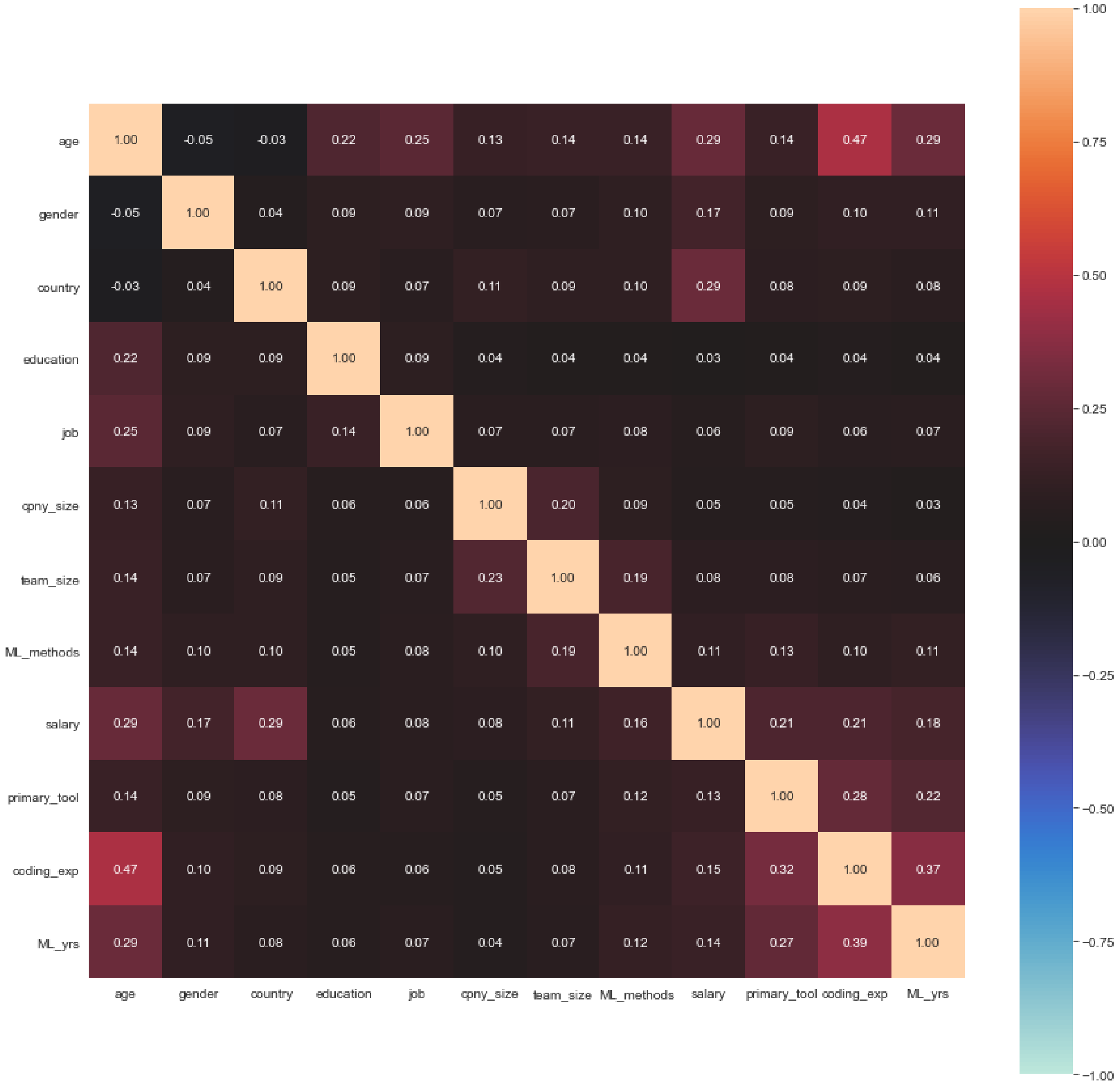


An even more significant difference lies in the last feature of our selection, the salary of data science workers. In the figure below, evidence shows that the salaries approximately follow a normal distribution (when excluding the outlier category of those who make $0-999 yearly income). Since we had to transform the salary variable into an ordinal variable, the reader should know that the distribution ranges from $0 to the last category of ">$500,000". While the difference between each category of income is not equal – the lowest categories have a $1000 range and the highest categories have variable ranges – some indicators can nonetheless provide some insight. For example, we find that the median salary among American data science workers is of $100,000 – 124,999. The median in Canada is much lower, with $60,000-69,999 worth of annual income.


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
25
salaryOrd
country
United States of America
Canada


Distribution of salaries
25
20
15
10
5
0
salaryOrd
United States of America
Canada
country

Lastly, we are able to find a statistical test for the association between all of our selected features, using *dython*, a library created by Shaked Zychlinski[1]



The figure above indicates that, among the features selected, few are strongly related to the country. Some of the rare strong associations found are not very informative (*e.g.* it goes without saying that age is related to salary and the number of years of experience using machine learning. The strongest correlation is with the salary, which confirms what we had observed above in the comparisons of distributions.

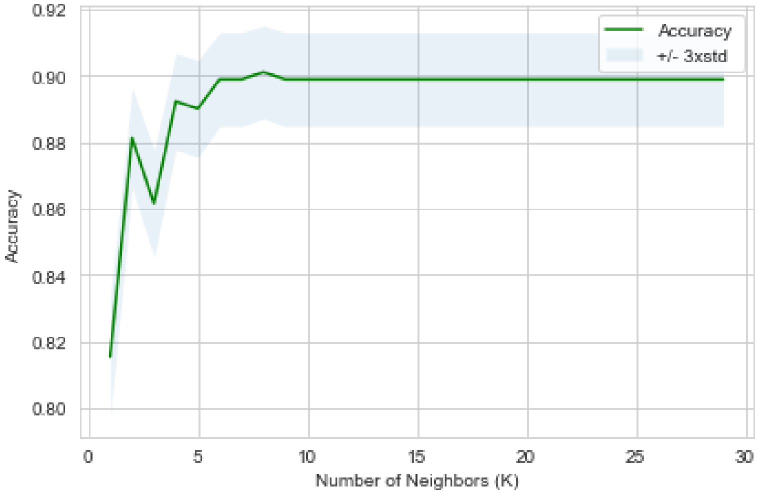
Modeling and Evaluation

At this stage, we prepare the data for modeling. We transform our features to ordinal variables when it makes sense to do so (*e.g.,* for age, salary, and number of years of experience with machine learning). Then, we convert nominal variables into dummy variables using the "one-hot encoding" technique. We drop the "I do not know" category of response. To deal with missing values, we simply drop the rows where there are missing values because the dataset is large enough already (2271 rows). We then normalize the features that will be used in our models. We conduct a train-test split for each model we will build. To make sure we select the best available model, we build the following four machine learning models in turn: logistic regression, K-nearest neighbour, decision tree, and support vector machine. The table below provides information to assess and compare the performance of our models.

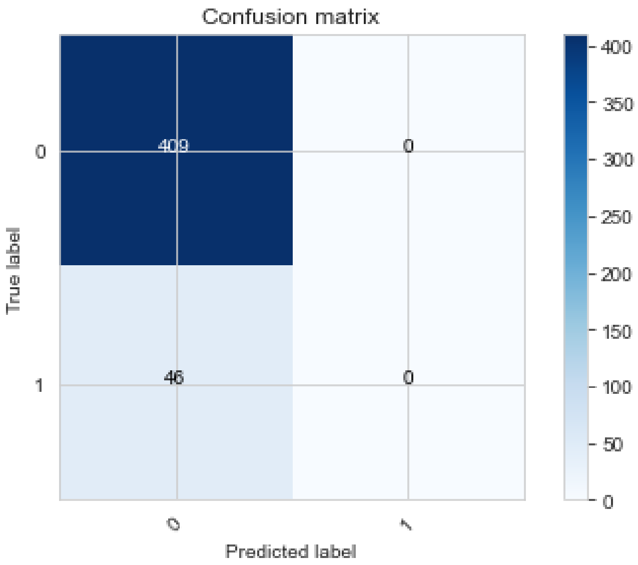
Model	**Accuracy** **(F-1, weighted average)**	**LogLoss**
Logistic Regression	0.8488	0.3216
K-Nearest Neighbour	0.8563	N/A
Decision Tree	0.8499	N/A
Support Vector Machine	0.8510	N/A

Overall, all the models perform quite well. There is not much of a difference between them, however. In the notebook, we notice that the maximum accuracy was reached by the

Decision Tree model (as shown in the figure below), with a score of over 90%. We seek to find out why the weighted averages shown in the table above are lower.



We detect a problem with predictions in the SVM model. In the confusion matrix shown below, we find that, while the model is excellent at predicting whether a respondent is American, it is unable to predict respondents who are Canadian. This effectively explains why the maximum accuracy drops when we look at the average weighted accuracy.



The figure above illustrates a recurrent problem across our models. By and large, the models do achieve excellent levels of accuracy on average, but they fail to predict 'Canadianness'.

Discussion & Recommendations

The exploratory data analysis and visualizations were useful in distinguishing between the Canadian and the American data science labour market. We now know that there is a significant wage gap between the two countries. We know that there are significant, but variable differences when it comes to job titles, company size, team size, years of experience with machine learning, level of machine learning maturity in the company, job title, primary tool used, education, and age. And we know that there is not much of a significant difference when it comes to the gender distribution. The field is overwhelmingly masculine in both countries.

While we were able to capture some differences between the two markets in the EDA, the models created failed to predict the Canadian identity of respondents. This might mean that, although we judged significant some of the differences observed in our visualizations, the models indicate a lack of statistical significance between the two markets.

All things considered, further research would be required to determine whether the differences observed were significant. An additional limitation precluding any definitive conclusion lies in the dataset we have used. All of the respondents were in some way "Kagglers", since, as Kaggle puts it, "most of [the] respondents were found primarily through Kaggle channels, like [their] email list, discussion forums and social media channels."[2] It would be interesting to know about data science professionals who are not part of the Kaggle community.

Conclusions: Future Deployment Opportunities

In conclusion, this study has provided several insights into the Canadian and American data science labour markets, and the differences between them. This might prove useful for recruiters seeking to better understanding the structure of the labour market in North America, for employers seeking to compare candidates and to compare their own company with the competition, as well as for prospective candidates who are beginning or transitioning towards a career in data science.

Future directions for research may include webscraping techniques to examine the features of current job postings over time. For instance, Andre Sionek, a data engineer, has already proposed a dataset he has scraped from Glassdoor in 2019[3], but companies need to remains cautious when it comes to the legal implications of webscraping techniques as this seems to be considered a "grey area".

Another possible area of inquiry would consist in assessing the evolution of data science labour markets over time. A good starting point would be to use the previous versions of the same Kaggle survey on data science and machine learning.

[1] Zychlinski, Shaked. 2020. "Dython." Online. http://shakedzy.xyz/dython/. Accessed November 3rd, 2020.

[2] Kaggle. 2019. "2019 Kaggle ML & DS Survey." Online. https://www.kaggle.com/c/kaggle-survey-2019/overview. Accessed November 3rd, 2020.

[3] Sionek, Andre. 2019. "Data Job Listings – Glassdoor." Online. https://www.kaggle.com/andresionek/data-jobs-listings-glassdoor. Accessed November 3rd, 2020.