Machine Learning Engineering Challenge

Name:

Sevda Ebrahimi

September 7, 2020

Table of Contents

Table of Contents 2
1. Introduction 3
2. Visualize Labour/Employment data 3
   2.1. Variation of labour and employment for the last 5 years-Both sexes 4
   2.2. Variation of labour force in the last 5 years-men and female in Ontario 6
   2.3. Comparison of the Variation of Employment force in the last 5 years-both sex (Canada, Ontario & Alberta) 7
3. Visualize a mashup correlating skills and jobs to industry and employment 8
4. References 11

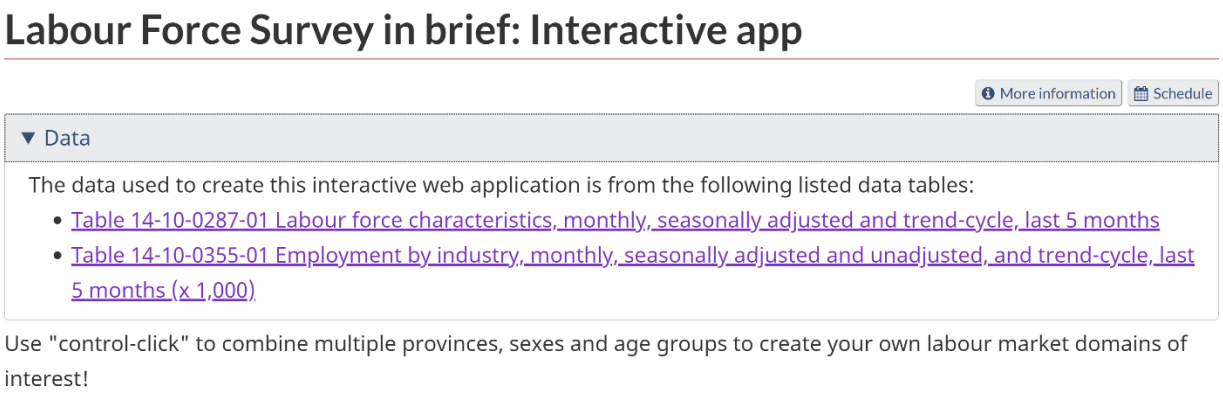
1. Introduction

In this report,

2. Visualize Labour/Employment data

In this section, the process of data visualizing will be explained. The datasets which have been used and the coding language for visualizing these datasets will be introduced.

There exist two datasets in the link referred.



First dataset, introduces labour force characteristics including:

- Population
- Number of labour force
- Number of Employment
- Number of Unemployment
- …

For Canada and all 11 province separately, for men, female and both sexes, for different month of different years from different ages.

I selected these three group to discuss and visualize, because the concept is the same for the other characters:

- Number of labour force
- Number of Employment
- Number of Unemployment

Considering all the age groups, dataset have different information based on sex and region for these characteristics. The information considered to be from April 2016 to August 2020(last 5 years).

The excel dataset, exist in the attached file with the name of '1410028701-eng'.

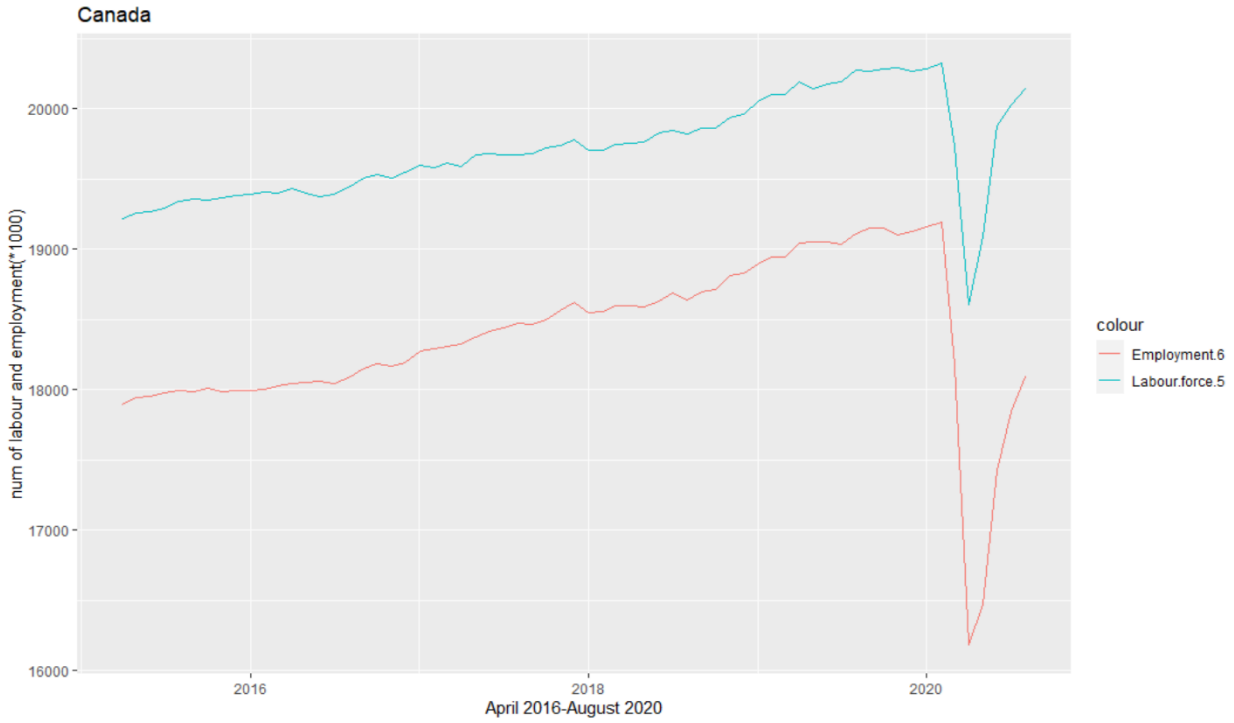
	A	B	C	D	E	F
1	Geography	sex	date	Labour force 5	Employment 6	Unemployment 7
2	Canada	Both sexes	4/1/2015	19213.4	17898.6	1314.8
3	Canada	Both sexes	5/1/2015	19258.9	17941.7	1317.1
4	Canada	Both sexes	6/1/2015	19269.3	17949	1320.3
5	Canada	Both sexes	7/1/2015	19294.5	17976.3	1318.2
6	Canada	Both sexes	8/1/2015	19342	17990.4	1351.6
7	Canada	Both sexes	9/1/2015	19354.7	17987.1	1367.7
8	Canada	Both sexes	10/1/2015	19353.2	18007.1	1346.1

R programming language is chosen for the Visualization of this dataset. The IDE for coding is RStudio too. This is the dataset structure demonstrated by str() command.

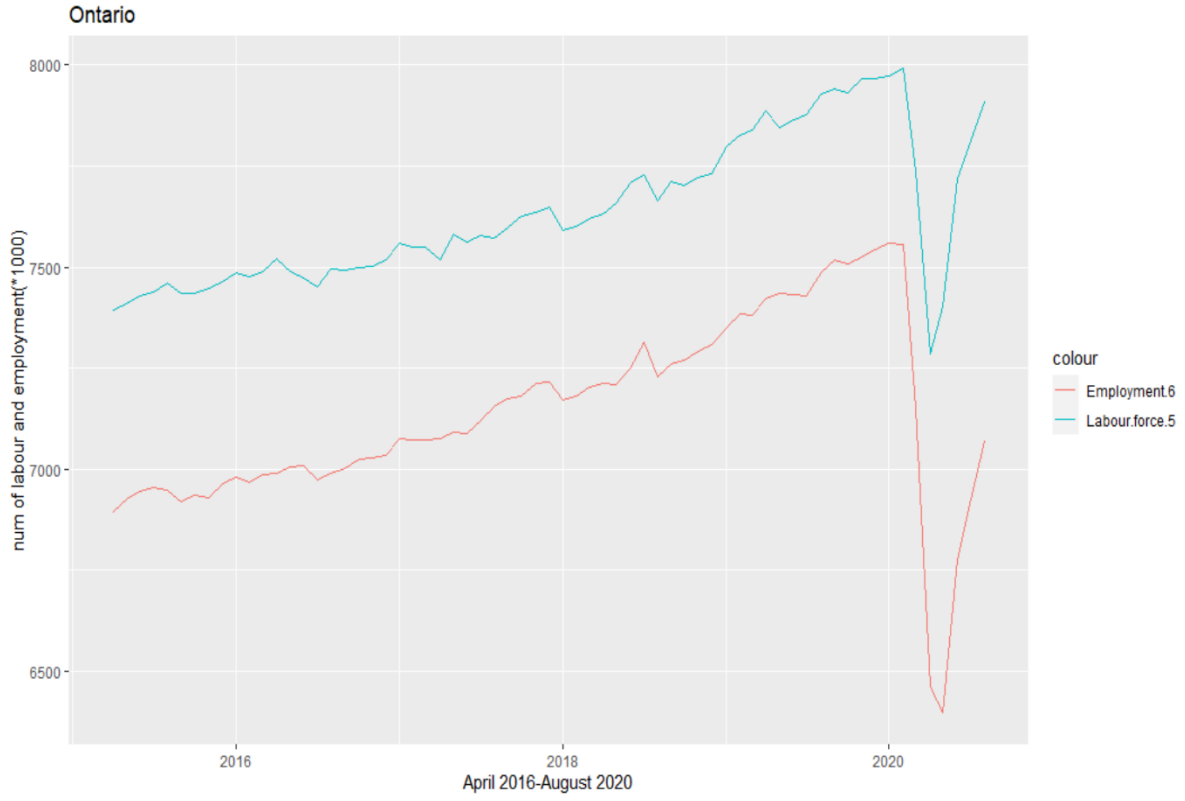
```
> str(df_new)
'data.frame':   585 obs. of  6 variables:
 $ Geography     : Factor w/ 3 levels "Alberta","Canada",..: 2 2 2 2 2 2
 ...
 $ sex           : Factor w/ 3 levels "Both sexes","Females",..: 1 1 1 1
 1 ...
 $ date          : Date, format: "2015-04-01" "2015-05-01" ...
 $ Labour.force.5: num  19213 19259 19269 19295 19342 ...
 $ Employment.6  : num  17899 17942 17949 17976 17990 ...
 $ Unemployment.7: num  1315 1317 1320 1318 1352 ...
>
```

2.1. Variation of labour and employment for the last 5 years-Both sexes

At this graph we could see the changes of labour force and employment for Canada, we clearly see the huge drop during Covid-19.

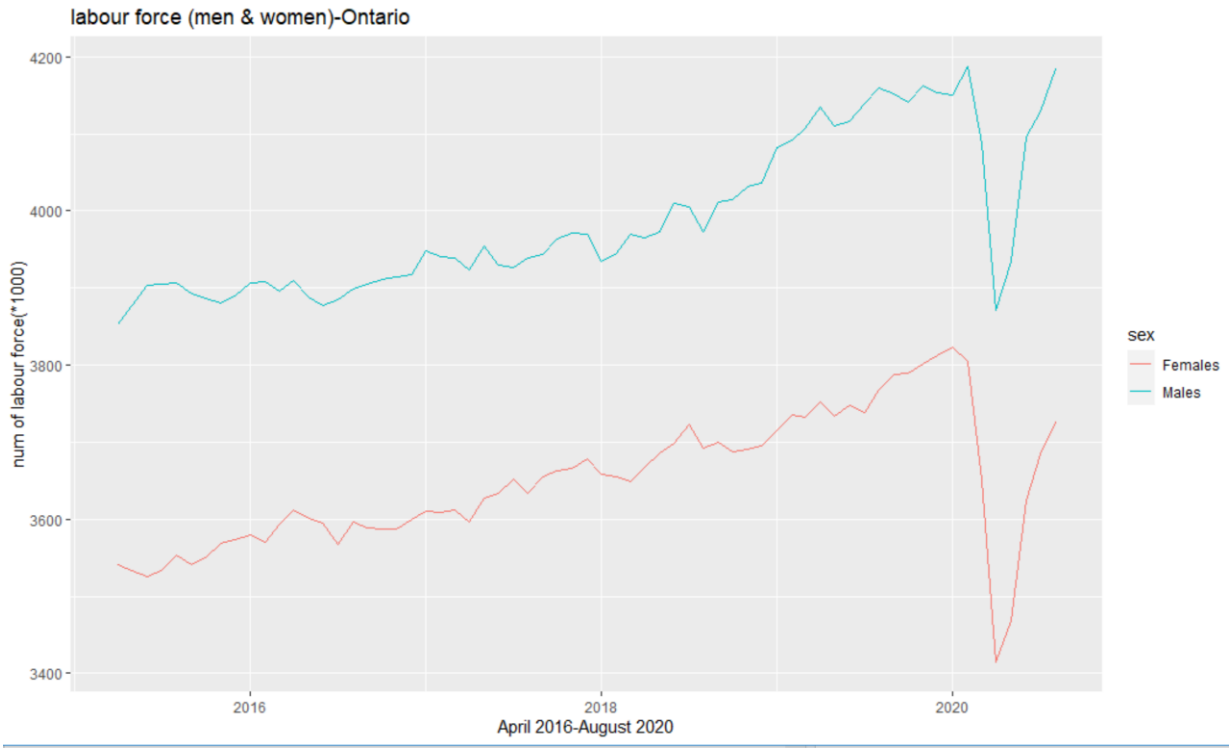


At this graph(below) we could see the changes of labour force and employment for Ontario , we clearly see the huge drop during Covid-19.



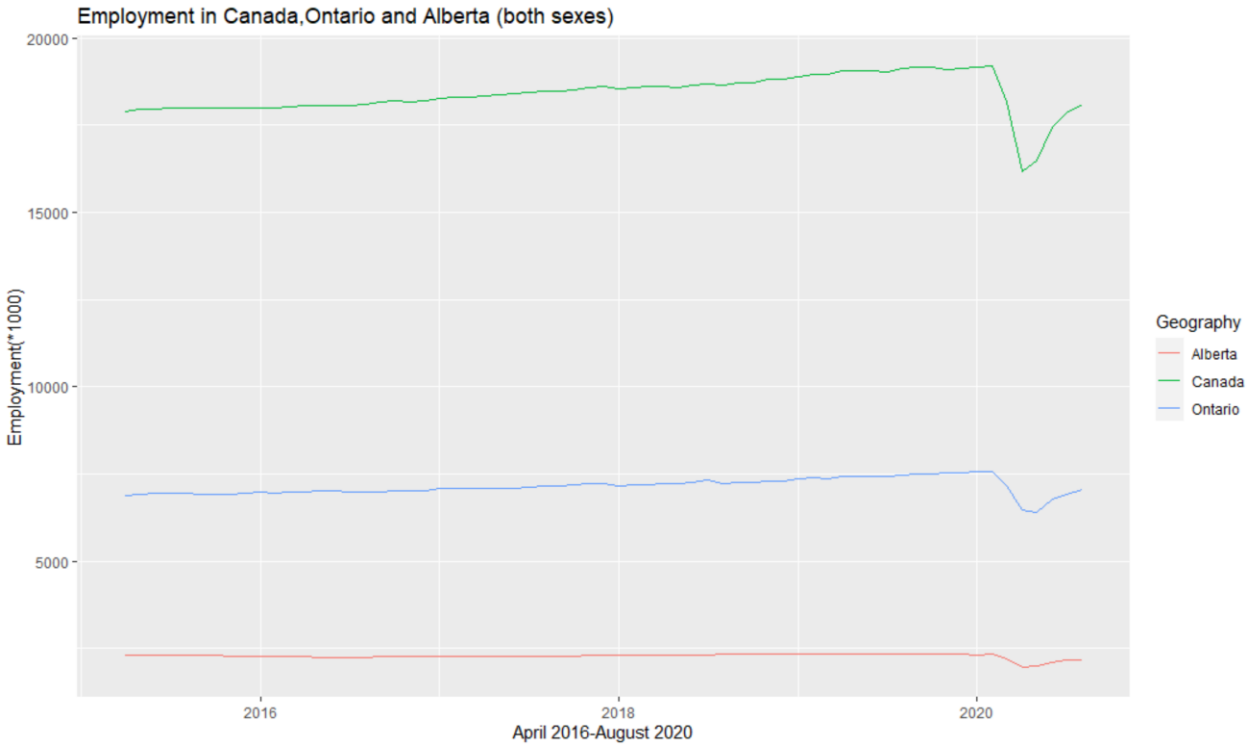
2.2. Variation of labour force in the last 5 years-men and female in Ontario

We could easily compare these two labour force in Ontario by this graph.



2.3. Comparison of the Variation of Employment force in the last 5 years- both sex (Canada, Ontario & Alberta)

This is the employment variation in Ontario and Alberta. There is three region in dataset, Canada, Ontario and Alberta. I choose Canada, Ontario and Alberta to comparison.



3. Visualize a mashup correlating skills and jobs to industry and employment

To do this task, there has to be something that enables us to find a correlation between these two dataset. These two datasets name are ‘1410035501-eng-industry’ and ‘job_skills’ in the attached file. A view of datasets are shown below.

	A	B	C	D	E	F	G	H
1	Company	Title	Category	Location	Responsibilities	Minimum Quali	Preferred Qualificati	
2	Google	Google Cloud Program M	Program Management	Singapore	Shape,	BA/BS degree	Experienc	
3	Google	Supplier Development En	Manufacturing & Supply Chain	Shanghai, China	Drive cross-	BS degree in an	BSEE,	
4	Google	Data Analyst, Product and	Technical Solutions	New York, NY, L	Collect and	Bachelorâ€™s	Experienc	
5	Google	Developer Advocate, Part	Developer Relations	Mountain View,	Work one-on-	BA/BS degree	Experienc	
6	Google	Program Manager, Audio	Program Management	Sunnyvale, CA, L	Plan	BA/BS degree	CTS	
7	Google	Associate Account Strateg	Technical Solutions	Dublin, Ireland	Communicate	Bachelor's	Experienc	
8	Google	Supplier Development En	Hardware Engineering	Mountain View,	Manage cross-	BS degree in	Master's	
9	Google	Strategic Technology Part	Partnerships	Sunnyvale, CA, L	Lead the	BA/BS degree	BA/BS	
10	Google	Manufacturing Business M	Manufacturing & Supply Chain	Xinyi District, Ta	Develop	BA/BS degree	MBA	
11	Google	Solutions Architect, Healt	Technical Solutions	New York, NY, L	Help compile	BA/BS degree	Master's	

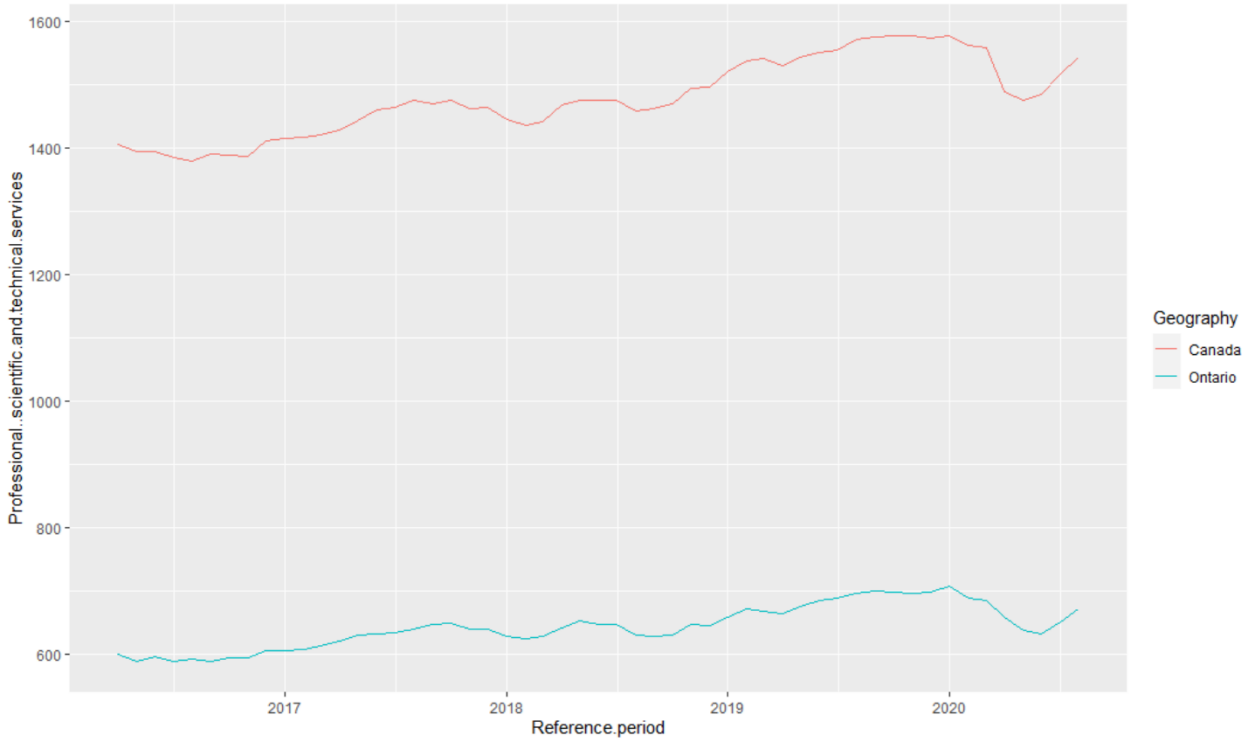
	A	B	C	D	E	F	G	H	I	J	K
1	Geography	Reference per	Agriculture 6	Forestry, fishi	Utilities	Constructio	Manufactu	Wholesale	Transporta	Finance, in	Professiona
2	Canada	4/1/2016	287.9	330.5	139.1	1384.6	1690.1	2775.3	905.6	1115.9	1405.2
3	Canada	5/1/2016	293	320.3	137.8	1401.3	1688.9	2729.6	900.9	1121.7	1394.4
4	Canada	6/1/2016	288.9	317.2	136.4	1376.7	1682	2745.3	906.1	1126	1394.6
5	Canada	7/1/2016	295.5	312.8	134.9	1378.3	1687	2727.8	912.3	1130.6	1384.8
6	Canada	8/1/2016	292.4	321.1	134.8	1386.8	1696.4	2745.6	910	1126.1	1378.9
7	Canada	9/1/2016	288.6	320.7	136.5	1387	1703.5	2734.7	917.7	1132.6	1391.3
8	Canada	10/1/2016	283.3	330	135.3	1401.8	1695.3	2752.5	917.9	1136	1389.5
9	Canada	11/1/2016	286.2	327.5	138.1	1384.4	1678	2750.3	912.2	1150.3	1387.7

If we analyse there two dataset we see there is not such a connection that we imagine to use machine learning algorithms. The first dataset shows some job titles which are branches of job categories. These dataset do not belong to Canada only, they are world wide and there is only 4 job category located in Canada(only Waterloo and Quebec). On the other hand we have second dataset which shows a different industries(finance, health, ...) employment information in different regions of Canada for the last 5 years and we want to find a correlation or correlation coefficient between two dataset. The other problem is that , we use dataset to train our Artificial Neural Network, to find the best and optimised weight of our network to do a specific task like image classification using CNNs or time series prediction using Recurrent Neural Networks and we do not train dataset. Dataset is some information which exist and we use dataset to train our models like what we do in computer vision or bioinformatics. But, I imagen that this job or skills distribution which we can extract from the job’s category in dataset are exactly the same of Canada, so the correlation coefficient is 1 between different categories of jobs in job_skills data and professional..scientific.and.technical.services column of ‘1410035501-eng-industry’ . I assume that all job category listed in job_skills are related to

professional..scientific.and.technical.services section of industry. So if we want to visualize each category of jobs in this section of industry, we have to multiply the probability of each category in existing dataset(jobs_skills) to the number of employment number during the last 5 years. Consequently we can see each category of job variation during the last 5 years.

here is the visualization and written codes in R:

First I have compared variation of professional, scientific and technical services section of industry in Ontario and Canada(*1000).



Then I find the frequency of each category of jobs in the whole categories using the job_skills dataset and assumed these distribution is as same as in Canada for professional, scientific and technical services section of industry (because all jobs were in google or youtube company I considered them to belong to this section).this the code for finding frequency and accordingly the probability of each category and then finally it's variation in Canada.

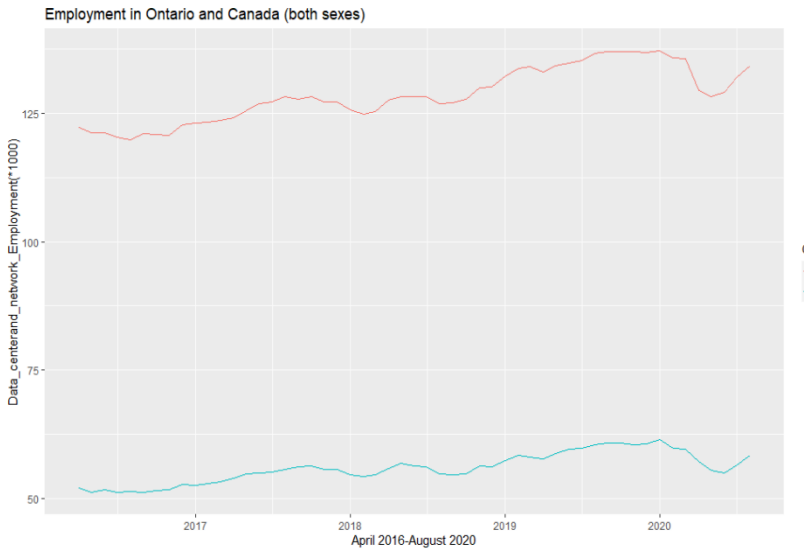
```
####jobs_skills data

jobs_skills<-read.csv('job_skills.csv')

str(jobs_skills)

jobs_skills$Category<-as.factor(jobs_skills$Category)
```

```
class(jobs_skills$Category)

str(jobs_skills)

levels(jobs_skills$Category)    ##23 type of job Category

Data_centerand_network_rate<-sum(jobs_skills$Category=="Data Center & Network")/23

dataframe_new<-
mutate(df_industry,Data_centerand_network_distributaion=df_industry$Professional..scientific.a
nd.technical.services*Data_centerand_network_rate)

ggplot(df_industry, aes(Reference.period, Data_centerand_network_distributaion, colour =
Geography)) +

  geom_line()
```

I selected the Data centerand network to show as an example, because there is 23 job categories.

```
> levels(jobs_skills$Category)##23 type of job Category
 [1] "Administrative"                    "Business Strategy"
 [3] "Data Center & Network"             "Developer Relations"
 [5] "Finance"                           "Hardware Engineering"
 [7] "IT & Data Management"              "Legal & Government Relations"
 [9] "Manufacturing & Supply Chain"      "Marketing & Communications"
[11] "Network Engineering"               "Partnerships"
[13] "People Operations"                 "Product & Customer Support"
[15] "Program Management"                "Real Estate & Workplace Services"
[17] "Sales & Account Management"        "Sales Operations"
```

And this the final visualization, showing the “Data Center and Network” variation in Canada for the last 5 years.



And if we want to use this information in machine leaning, we can use these information as a time series to predict the future distribution of job positions and the country can plan for it's future. We can use deep learning for forecasting. We could use multilayer perceptron NNs, Recurrent NNs, CNNs for this aim. This time series can be framed supervised learning problem and regression because we do not want to classify something, we just want to predict number of possible employment for each category of job.

4. References

[1] https://www150.statcan.gc.ca/n1/pub/14-20-0001/142000012018001-eng.htm