Final Project Report

CMPT 353 Computational Data Analysis, Summer 2023

Isaac Ding 301425524

Huanyu Zhou 301417467

Kaize Gu 301416566

Table of Contents

1.Introduction

1.1 Summary of Project

The COVID-19 pandemic that started in 2019 was arguably one of the most significant events that happened in the recent years. Causing more than 760,000,000 cases and nearly 7,000,000 deaths across the globe (WHO Coronavirus,2019), The pandemic has caused much more consequences than just the diseases: entire countries were under quarantine, international trades were hard to conduct, economy was damaged, and households went bankrupt.

Among all the side effects of the pandemic, one of the most obvious and well-recorded one was the unemployment rate. This project aims to use statistical tests, linear and polynomial regression as well as machine learning techniques to understand how the covid cases and deaths in Canada correlate to the unemployment of Canada.

1.2 Data Used

Since we are studying the influence of Canada's COVID-19 pandemic on Canada's unemployment rate, We took data from the following two tables:

"Labour force characteristics by province, monthly, seasonally adjusted", retrieved from Statistics Canada, and "Public Health Infobase - Data on COVID-19 in Canada", retrieved from Government of Canada.

The first table provides the monthly employment-related data seasonally adjusted for both sexes and all provinces/territories of Canada, and the second table provides weekly covid cases and covid deaths for all provinces/territories of Canada. With these two tables we are able to obtain all information related to our project.

1.3 Problems to Solve

With the data collected above, we aim to provide questions for the following groups of questions:

1. Does the unemployment rate perform differently for different genders and different regions in Canada? Does the COVID-19 pandemic relate to the unemployment rate?
2. If the unemployment rate is related to the COVID-19 pandemic, what relationship is between the unemployment rate and the COVID-19 cases/deaths? Is the covid cases/deaths linearly related to the unemployment rate. How much does the COVID-19 cases/deaths affect the unemployment rate of each province/territory?
3. Is it possible to train machine learning models to predict a region's unemployment solely based on the covid cases/deaths? Is it possible to infer about a region's COVID-19 cases or deaths using the employment statistics of the region?

Question group 1 will be answered in the Data Gathering & Cleaning and Statistical Tests sections. Question group 2 will be answered in the Data Preprocessing and Multilinear/Polynomial Regression sections. Question group 3 will be answered in the Machine Learning section.

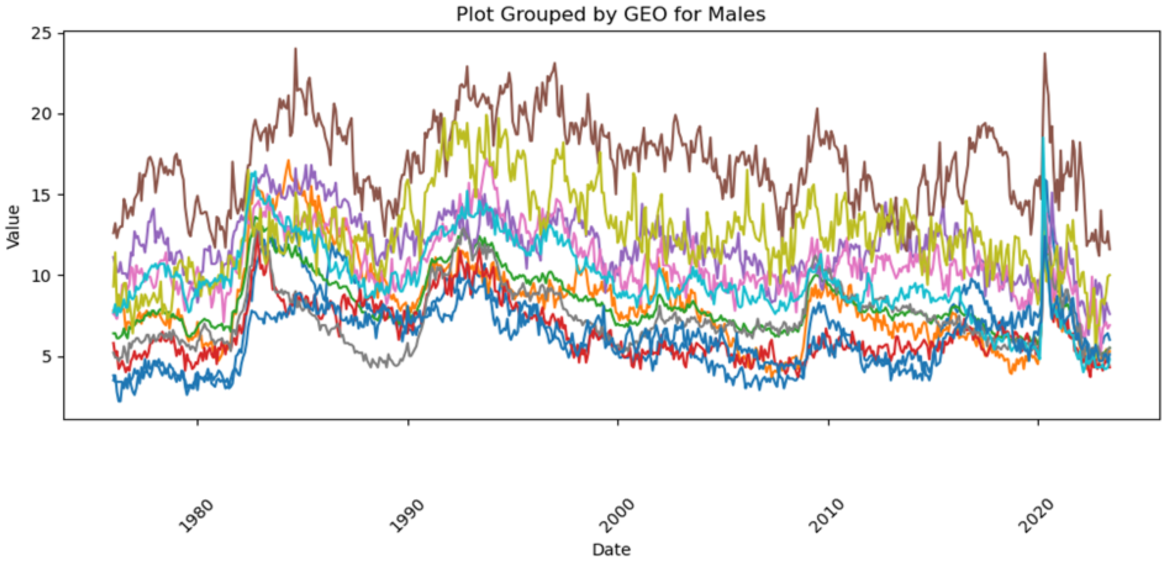
2. Data Gathering & Cleaning

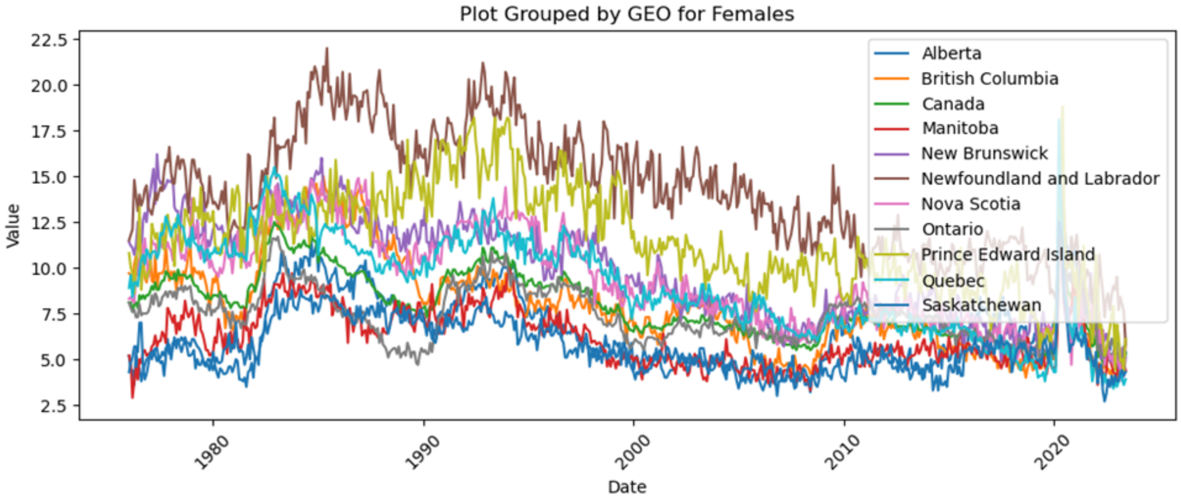
In this project, we are focusing on the relationship between unemployment rate data and covid-19 case data. However, the Labour force statistics provided by Statistics Canada is too chaotic to provide the data we need, so we did the following data cleaning:

The combined data does not use the table format we usually see, but a “characteristic-value” pair format. For example, if the ‘Labour force characteristics’ = ‘Unemployment rate’, ‘GEO’ = ‘British Columbia’, ‘Sex’ = ‘Male’, ‘VALUE’ =’4.9’, and ‘REF_DATE’= ‘Apr-19’ that means in British Columbia, the unemployment rate for male in 2019-Apr is 4.9%. After filtering the data to ‘Labour force characteristics’ = ‘Unemployment rate’, ‘Data type' = 'Seasonally adjusted' and 'Statistics' = 'Estimate', we are able to focus on the unemployment rate data and test our hypothesis. Then I plot the unemployment rate over time for different regions in Canada for both males and females to see if there is any useful observation. The code is provided in the ‘Clean&Visualization.py’ file.

3. Data Processing & Statistical Tests

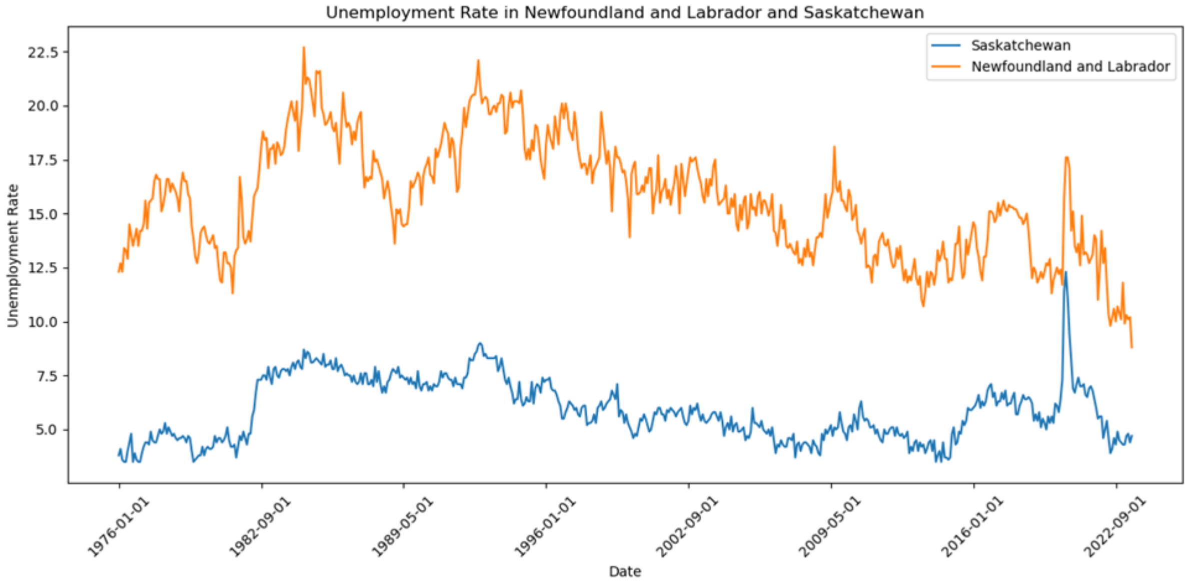
3.1 Visualizing Unemployment rate from 1976-2023





We can see the region of ‘Newfoundland and Labrador’ has a relatively higher unemployment rate than Saskatchewan, it comes to a hypothesis that there is a difference between the unemployment rate data in these two provinces. To achieve this we drop the data from all the other regions and keep these two provinces for comparison.

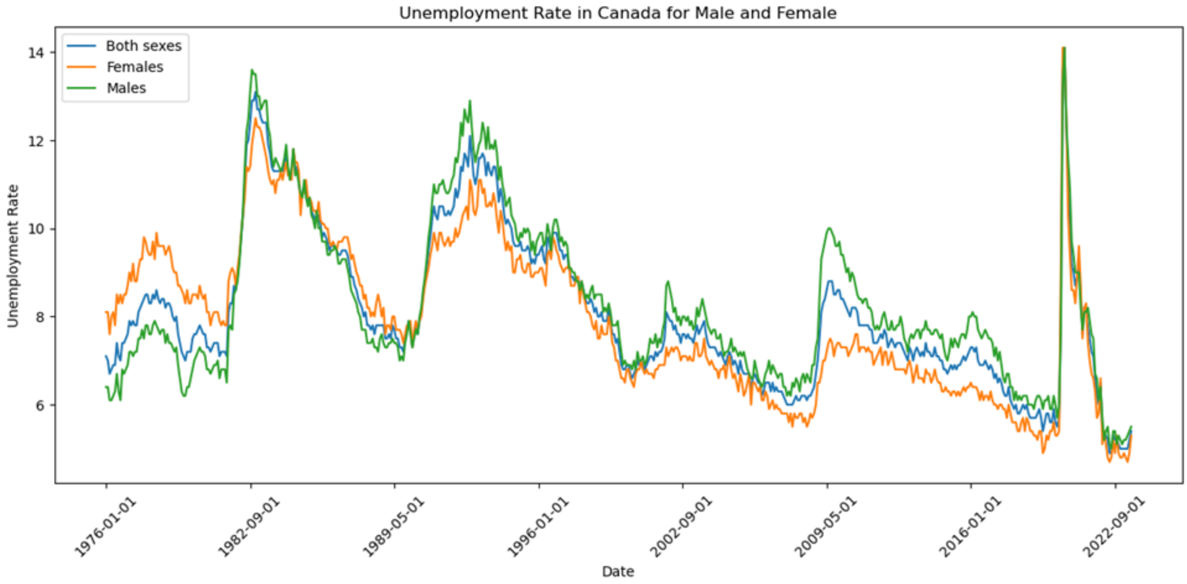
3.2 Visualizing Unemployment in Two Regions for Both Sexes



As shown in the plot, the unemployment rate in 'Newfoundland and Labrador' is always higher than that in 'Saskatchewan', however, to further confirm our hypothesis, we applied the Mnn-Whitney U test on it and get the output p_value = 1.3514187856105966e-187. As the p_value < 0.05, we can reject the null hypothesis, and conclude there is a significant difference between the data in these two regions. The code is provided in the ‘GEO_comparision.py’ file.

After that, we would like to test if there’s a difference between the unemployment rate for males and females, as the data visualization for 11 different regions is too chaotic to draw a valid conclusion, we will investigate the data in the region of whole Canada this time, and plot it again.

3.3 Visualizing Unemployment Rate for Different Sexes

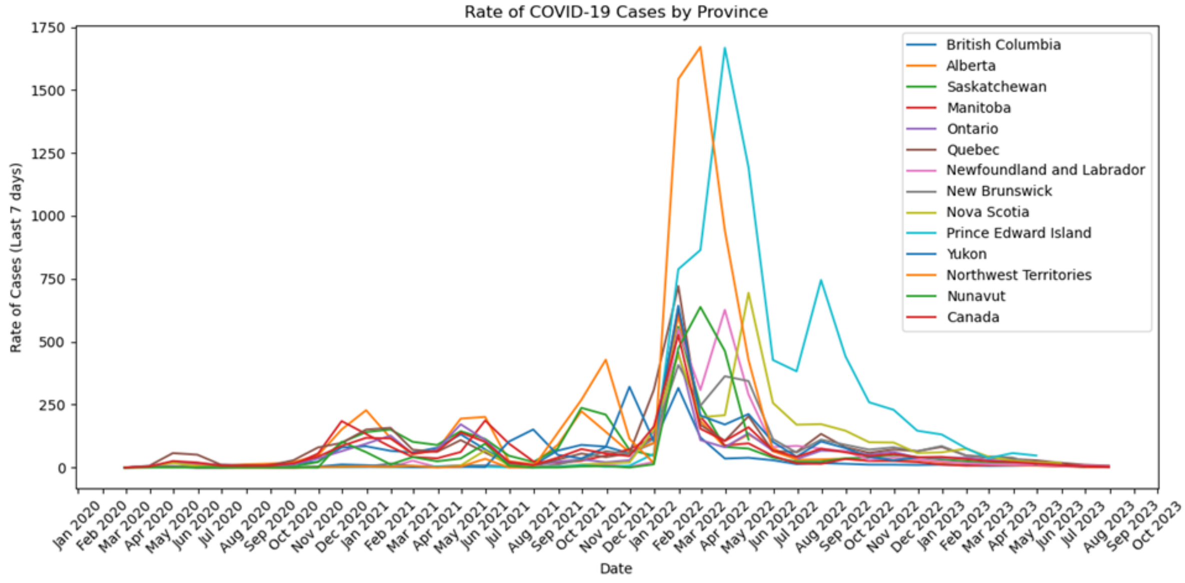


In addition to males and females, we also visualized the plot for ‘Both sexes’. It can be seen that as the male and female unemployment rate fluctuates, they follow very similar trends, and are both very close to the unemployment rates for both sexes. We think it is reasonable to represent male and female unemployment rate by the unemployment rate of both sexes, and it will help make more useful results.

The code is provided in the ‘Unempl_Sex_Analysis.py’ file.

Before we start analyzing the relationship between the selected two datasets, we first analyze the covid-19 average case rate.

3.4 Visualizing Covid-19 Weekly Cases



From the plot we can see the case rate for all the provinces significantly increased in Dec 2021 and reached a peak at around Feb 2022 to Apr 2022.

As the data from Alberta seems to have a relatively larger statistic than other provinces, we did a hypothesis test to test if there's a significant difference between data in Alberta and a random province we choose, which is Yukon.

Output:

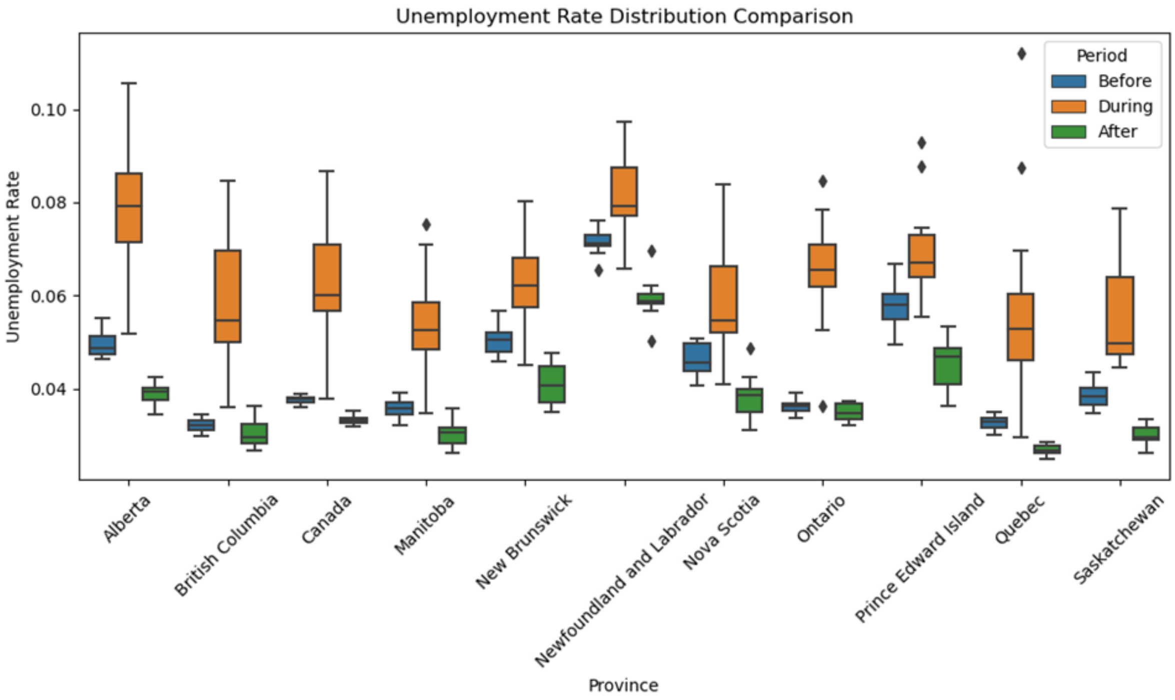
p_value = 0.9955263219772348

We fail to reject the null hypothesis. There is no significant difference between the data in Alberta and Yukon.

To compare the unemployment rate in different time periods, we split our unemployment data into three subsets, Jan 2019 – Jan 2020 represents data before the pandemic, Feb 2020 – Feb 2021 represents data during the pandemic, and June 2022 – June 2023 represents data after the pandemic.

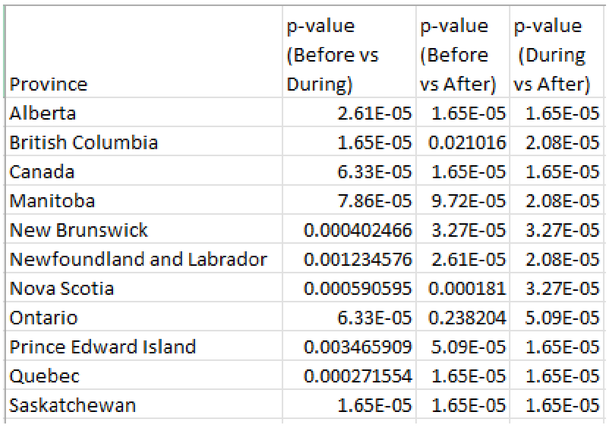
We first create a box plot for data visualization.

3.5 Unemployment rate comparison among Before, During and After the Pandemic



As shown in the plot, the unemployment rate data during the pandemic is much higher than that before the pandemic and after the pandemic in all the provinces in Canada. To further confirm our hypothesis, we did the Mann-Whitney U test among each indicator from each region, the p_value result was shown in the table below:

3.6 Unemployment rate comparison p-value in different provinces:



Province	p-value (Before vs During)	p-value (Before vs After)	p-value (During vs After)
Alberta	2.61E-05	1.65E-05	1.65E-05
British Columbia	1.65E-05	0.021016	2.08E-05
Canada	6.33E-05	1.65E-05	1.65E-05
Manitoba	7.86E-05	9.72E-05	2.08E-05
New Brunswick	0.000402466	3.27E-05	3.27E-05
Newfoundland and Labrador	0.001234576	2.61E-05	2.08E-05
Nova Scotia	0.000590595	0.000181	3.27E-05
Ontario	6.33E-05	0.238204	5.09E-05
Prince Edward Island	0.003465909	5.09E-05	1.65E-05
Quebec	0.000271554	1.65E-05	1.65E-05
Saskatchewan	1.65E-05	1.65E-05	1.65E-05

In the p-value table, except the comparison between unemployment rate data in Ontario before and after the pandemic has a p-value = 0.238204, all the other Mann-Whitney U tests have a p-value smaller than 0.05. So we can conclude, in all the regions we are investigating, the unemployment rate has a significant difference between the data before vs. during, before vs. after, and during vs. after the pandemic except for the data of Ontario before and after the pandemic. And we can conclude in general the pandemic has a significant effect on the unemployment rate.

The code is provided in the 'Before_After_Comparision.py' file.

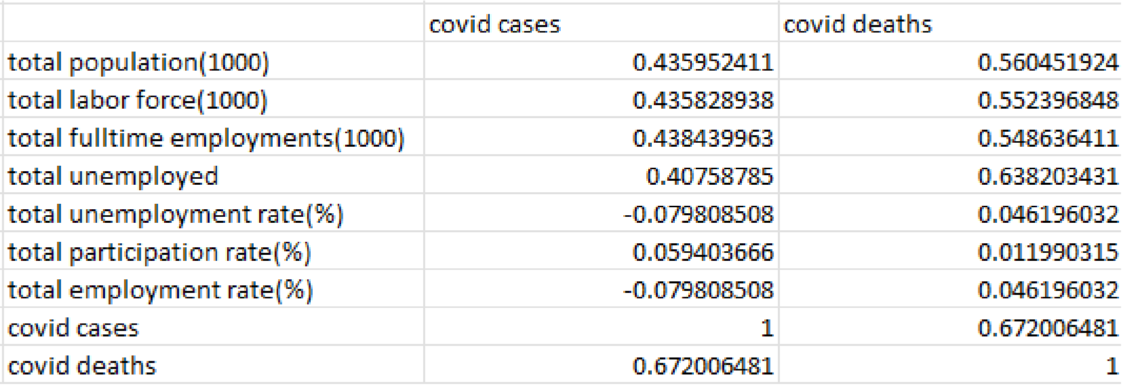
4. Multilinear/Polynomial Regression

4.1 Linear Correlation Test

With establishing that the COVID-19 pandemic does have an effect on the unemployment rate across Canada, the next problem becomes what is the relationship between the pandemic and the unemployment rate, and which provinces/territories are most affected by the covid cases/deaths in Canada. And since we concluded that the male and female unemployment rate share the same trend, we will use the unemployment data for both sexes from here on.

The first thing to do is to determine whether a region's COVID-19 cases and deaths are linearly correlated to the unemployment rate in the regions. We performed pearson's correlation tests, and it seems that the linear correlation is rather weak:(original chart can be found in covid_employment_correlations.csv)

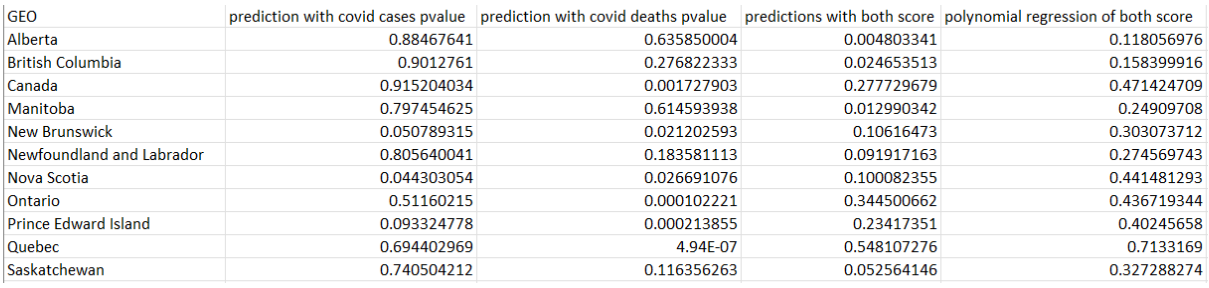
The unemployment rate has pearson's correlation value of about -0.08 with the amount of covid cases, and about 0.05 with the amount of covid deaths. This means there is a very weak linear relation between these two covid data and the unemployment rate.



	covid cases	covid deaths
total population(1000)	0.435952411	0.560451924
total labor force(1000)	0.435828938	0.552396848
total fulltime employments(1000)	0.438439963	0.548636411
total unemployed	0.40758785	0.638203431
total unemployment rate(%)	-0.079808508	0.046196032
total participation rate(%)	0.059403666	0.011990315
total employment rate(%)	-0.079808508	0.046196032
covid cases	1	0.672006481
covid deaths	0.672006481	1

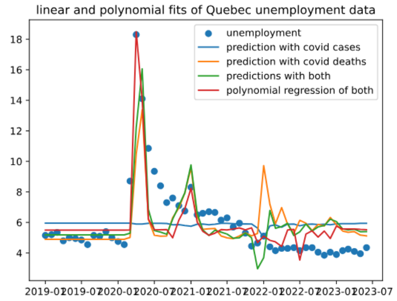
4.2 Trial 1: Fitting a Region's COVID-19 Data and Unemployment Rate

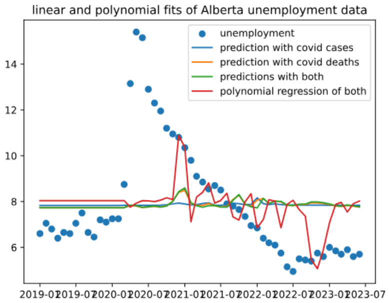
We therefore decided that maybe adding polynomial features as well can help with providing more flexibility with the model. Here is the model's performance with using the covid cases and deaths of a region as X, and the unemployment rate of that region as y:

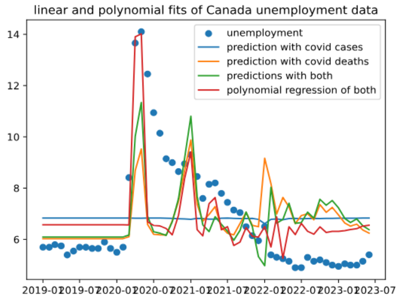


GEO	prediction with covid cases pvalue	prediction with covid deaths pvalue	predictions with both score	polynomial regression of both score
Alberta	0.88467641	0.635850004	0.004803341	0.118056976
British Columbia	0.9012761	0.276822333	0.024653513	0.158399916
Canada	0.915204034	0.001727903	0.277729679	0.471424709
Manitoba	0.797454625	0.614593938	0.012990342	0.24909708
New Brunswick	0.050789315	0.021202593	0.10616473	0.303073712
Newfoundland and Labrador	0.805640041	0.183581113	0.091917163	0.274569743
Nova Scotia	0.044303054	0.026691076	0.100082355	0.441481293
Ontario	0.51160215	0.000102221	0.344500662	0.436719344
Prince Edward Island	0.093324778	0.000213855	0.23417351	0.40245658
Quebec	0.694402969	4.94E-07	0.548107276	0.7133169
Saskatchewan	0.740504212	0.116356263	0.052564146	0.327288274

We can see performing linear regression with covid cases all failed to reject the null hypothesis, performing linear regression with covid deaths was better, but still failed to pass the test often. For columns 3 and 4 we used sklearn models and therefore we measured the training score instead of the p-value. In general, all models didn't perform very well. Here is a comparison of the best accuracy region, worst accuracy region, and the entire Canada's prediction visualization:

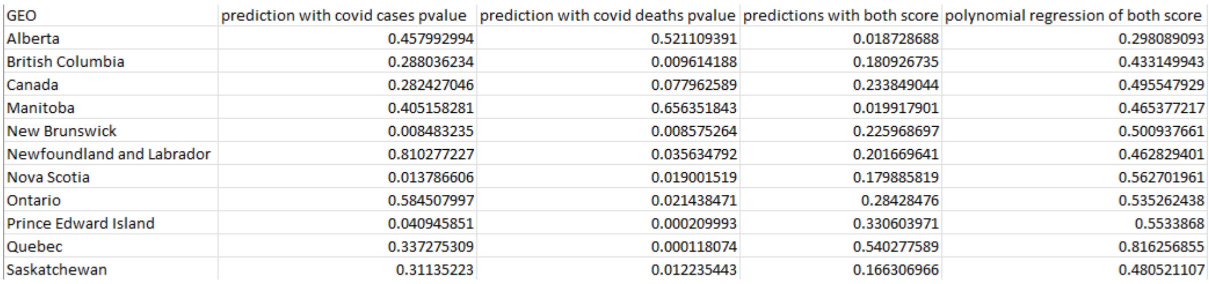






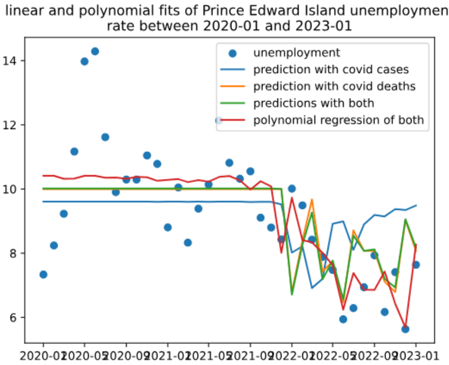
4.3 Trial 2: Fitting Only Data Where There Are More COVID-19 Cases

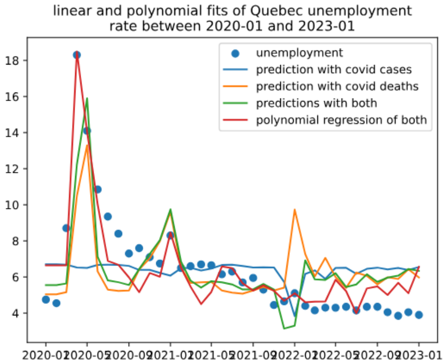
After doing the first linear/polynomial regressions, we realized that there are zero or very few covid cases at the time intervals before 2020-01 and after 2023-01, and thus the linear fit models can only fit to it by adjusting the intercept term, which does not provide the model enough flexibility. We also predicted that maybe because these time frames had little or no covid cases, they actually exhibit the pattern of the economy without the pandemic's influence, and maybe removing these data points can help the model to better predict the unemployment while it is under the pandemic's influence. This is the models' performance after removing the data points from before 2020-01 and after 2022-01:

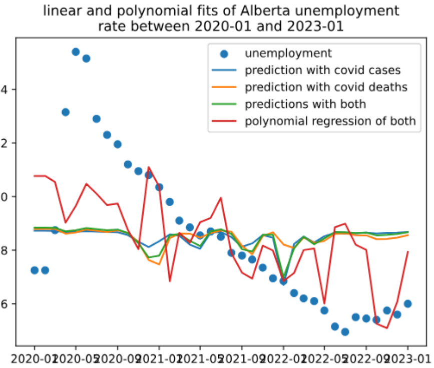


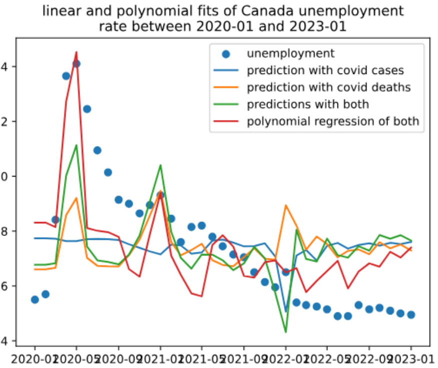
GEO	prediction with covid cases pvalue	prediction with covid deaths pvalue	predictions with both score	polynomial regression of both score
Alberta	0.457992994	0.521109391	0.018728688	0.298089093
British Columbia	0.288036234	0.009614188	0.180926735	0.433149943
Canada	0.282427046	0.077962589	0.233849044	0.495547929
Manitoba	0.405158281	0.656351843	0.019917901	0.465377217
New Brunswick	0.008483235	0.008575264	0.225968697	0.500937661
Newfoundland and Labrador	0.810277227	0.035634792	0.201669641	0.462829401
Nova Scotia	0.013786606	0.019001519	0.179885819	0.562701961
Ontario	0.584507997	0.021438471	0.28428476	0.535262438
Prince Edward Island	0.040945851	0.000209993	0.330603971	0.5533868
Quebec	0.337275309	0.000118074	0.540277589	0.816256855
Saskatchewan	0.31135223	0.012235443	0.166306966	0.480521107

We can see that the p-values for the first two linear regression improved a lot. In the second column, 8 regions are able to reject the null hypothesis instead of 6 from the last trial. Both multilinear regression(column 3) and polynomial regression(column 4) also improved the score in general. For the first two columns, the best predicted results came from Prince Edward Island, and for the later two columns, Quebec had the highest score. The models still performed worst in Alberta. Here is the predictions made for these regions by the models, along with the prediction made for the entire Canada:





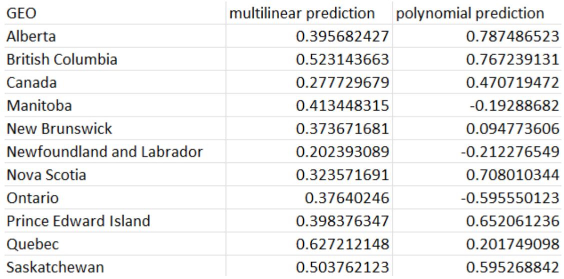




We can see that for Edward Island where both linear regressions had very good p-values, the predictions aren't really good, Whereas for Quebec where the multilinear and polynomial regressions had the best score, the predictions by these models closely follow the actual unemployment rate. This shows us that p-value isn't a great measure of performance in terms of linear regression model, and since linear regression with a single input is not very flexible, we decided to only use multi-linear and polynomial regression for the last trial. Aside from removing two models for evaluation, we are able to improve the performance of our prediction by constraining the time frame to when the covid cases and deaths numbers are more significant. This gives us the conclusion that the unemployment rate exhibited different patterns when the covid cases and deaths are more significant in Canada.

4.4 Trial 3: fitting with local and entire Canada COVID-19 data

If we look at Alberta, where our models performed worst, we can see that although Alberta had much different COVID-19 cases and deaths than the average of Canada (section 3.4), it still had the unemployment rate trend similar to the entire Canada's unemployment rate (section 4.3). This hints us that regions are affected not only by the local COVID-19 cases and deaths of that region, but also the cases and deaths of the entire country. With this we decided to add the entire Canada’s covid cases and deaths as features in addition to the region’s covid cases and deaths, when predicting a region’s unemployment rate. This gave us the following results:



GEO	multilinear prediction	polynomial prediction
Alberta	0.395682427	0.787486523
British Columbia	0.523143663	0.767239131
Canada	0.277729679	0.470719472
Manitoba	0.413448315	-0.19288682
New Brunswick	0.373671681	0.094773606
Newfoundland and Labrador	0.202393089	-0.212276549
Nova Scotia	0.323571691	0.708010344
Ontario	0.37640246	-0.595550123
Prince Edward Island	0.398376347	0.652061236
Quebec	0.627212148	0.201749098
Saskatchewan	0.503762123	0.595268842

We can see that for the multilinear prediction, most regions improved in accuracy compared to before augmenting in the additional entire Canada’s covid cases and deaths (Trial 1), whereas the accuracy of the “Canada” row stayed the same, as expected. This proves our point that regions are affected by the entire country’s pandemic as well as the local pandemic situation.

However, when we look at the polynomial prediction, some regions had a drastic boost in accuracy (Alberta’s accuracy went from 0.12 to 0.79 compared to before augmenting the Canada COVID-19 data), some went down a lot, and even went to negative values(e.g. Manitoba). This is likely because four input features gave the polynomial regression model too much flexibility, that our amount of data was unable to sufficiently train it.

4.5 Interpreting the Accuracy

We have tried many ways to improve the accuracy of our regression models, but it is also important that we know what these accuracy numbers can mean. When we perform linear, multilinear, or polynomial regressions of input features X to output features y, we are actually finding coefficients to a linear combination or polynomial of features in X that is as close to the outputs y as possible. The training accuracy measures how close are the coefficients to matching the actual output.

This means if we are able to find coefficients that give our model a 90% accuracy, the linear combination or polynomial of the features can predict the actual values of y with a 90% accuracy. In other words, a linear combination or polynomial of our input features X are able to explain 90% of the variances in the output feature y.

The best we got from our models is in our multilinear regression model for Trial 3 (Section 4.4), where our accuracy ranged from 0.20 to 0.63. This means a linear combination of local COVID-19 cases and deaths and countrywide COVID-19 cases and deaths can explain 20% to 63% of the unemployment rate, depending on different regions.

We can also think that if our model can explain less of the unemployment changes of a region, that the region is less directly affected by the COVID-19 cases and deaths, and vice versa. In this way of thinking, Newfoundland and Labrador’s unemployment rate is least directly affected by the COVID-19 cases and deaths in Canada, and Quebec’s unemployment rate is most directly affected by the COVID-19 cases and deaths in Canada.

5. Machine Learning

We can leverage machine learning methods to make predictions on the given data, specifically employing two commonly used techniques — Random Forests (RF) and k-Nearest Neighbors (k-NN). In order to evaluate the performance of these models, we calculate and compare two critical metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE, which quantifies the expected value of the square of the difference between predicted and actual values, is sensitive to outliers, while MAE, averaging the absolute values of the difference between predicted and actual values, shows less sensitivity to outliers. Comparing these two metrics enables us to gain a more comprehensive understanding of the model's performance.

In addition, assessing feature importance is crucial as it provides insights into the relative importance of individual features in model predictions. For instance, within the RF model, the significance of COVID-19 cases and deaths for predictions can differ depending on the city.

In Alberta, it has been observed that the number of cases influences unemployment prediction more significantly than the number of deaths. This indicates the necessity of considering the specific context of the problem when deciding which features to incorporate into the model.

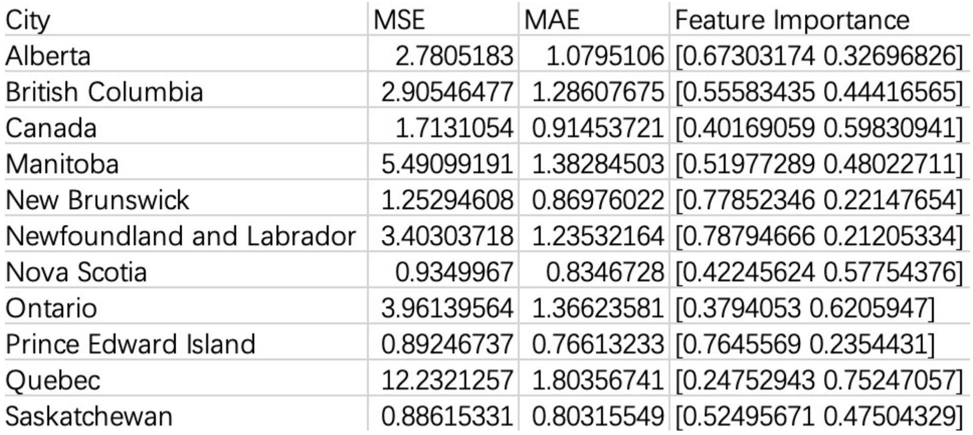
We use COVID-19 case, death, and unemployment rate data for various cities in Canada. The features include 'covid cases' and 'covid deaths' and the target variable is 'total unemployment rate(%)'.
We used the following two machine learning methods:

1. Random Forest (RF): RF is an ensemble learning method based on decision trees. We set the parameters as: number of trees = 100, minimum samples for a split = 2, minimum samples per leaf = 1.

2. k-Nearest Neighbors (k-NN): k-NN is an instance-based learning or a local approximation and simplification method. We set k = 5 and standardized the features.

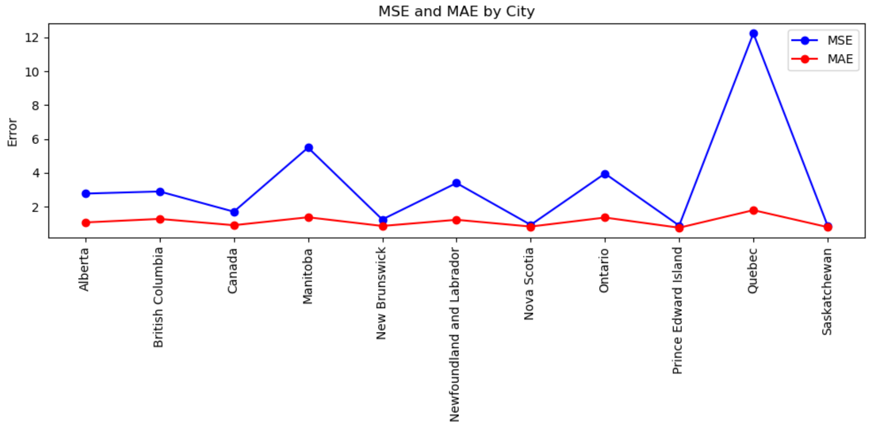
Here are our predictive results for various cities in Canada:

1. Random Forest: Across the cities, the MSE values vary between 0.89 (Prince Edward Island) to 12.23 (Quebec), and MAE values range from 0.83 (Nova Scotia) to 1.80(Quebec).



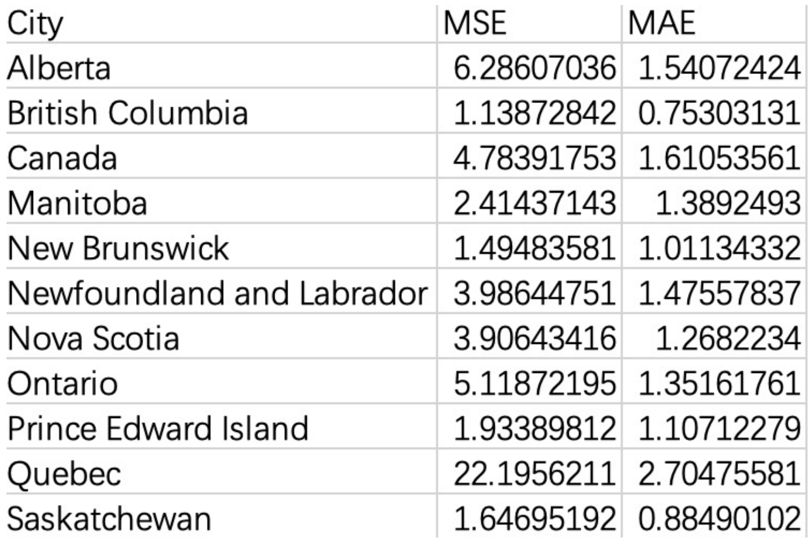
City	MSE	MAE	Feature Importance
Alberta	2.7805183	1.0795106	[0.67303174 0.32696826]
British Columbia	2.90546477	1.28607675	[0.55583435 0.44416565]
Canada	1.7131054	0.91453721	[0.40169059 0.59830941]
Manitoba	5.49099191	1.38284503	[0.51977289 0.48022711]
New Brunswick	1.25294608	0.86976022	[0.77852346 0.22147654]
Newfoundland and Labrador	3.40303718	1.23532164	[0.78794666 0.21205334]
Nova Scotia	0.9349967	0.8346728	[0.42245624 0.57754376]
Ontario	3.96139564	1.36623581	[0.3794053 0.6205947]
Prince Edward Island	0.89246737	0.76613233	[0.7645569 0.2354431]
Quebec	12.2321257	1.80356741	[0.24752943 0.75247057]
Saskatchewan	0.88615331	0.80315549	[0.52495671 0.47504329]

For instance, Alberta has an MSE of 2.780518299 and an MAE of 1.079510596. The importance of 'covid cases' and 'covid deaths' in the prediction model for Alberta are 0.67303174 and 0.32696826 respectively, suggesting that the number of COVID-19 cases is still more influential than the number of deaths when predicting the unemployment rate in this city. It's important to note that these numbers can vary depending on changes in the underlying data or alterations to the parameters of the machine learning model.



The plot graph I provided above can display the trend of data changes over time or a relative comparison between the data. According to these data, the model performance of Prince Edward Island and Saskatchewan is the best, as they have the lowest MSE and MAE values.

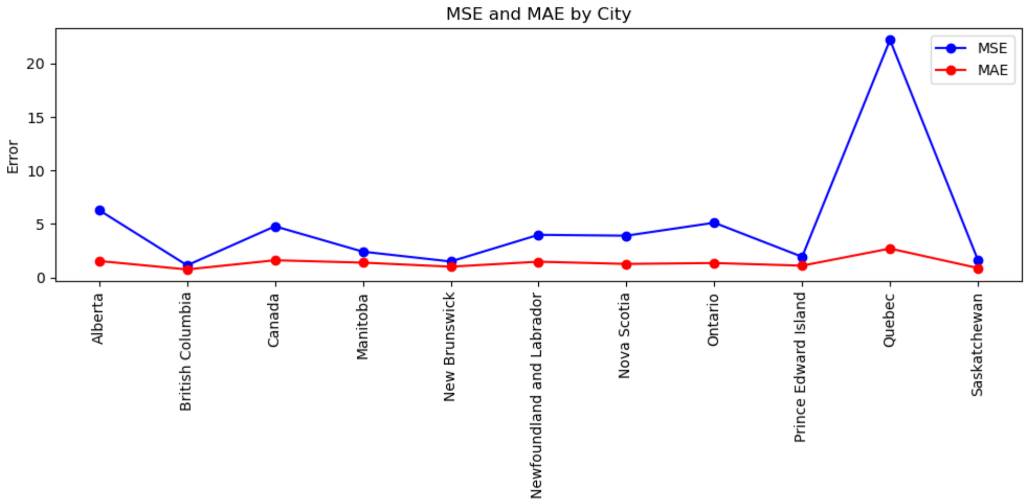
2. k-Nearest Neighbors: Across the cities, the MSE values vary between 1.14 (British Columbia) to 22.19 (Quebec), and MAE values range from 0.75 (British Columbia) to 2.70(Quebec).



City	MSE	MAE
Alberta	6.28607036	1.54072424
British Columbia	1.13872842	0.75303131
Canada	4.78391753	1.61053561
Manitoba	2.41437143	1.3892493
New Brunswick	1.49483581	1.01134332
Newfoundland and Labrador	3.98644751	1.47557837
Nova Scotia	3.90643416	1.2682234
Ontario	5.11872195	1.35161761
Prince Edward Island	1.93389812	1.10712279
Quebec	22.1956211	2.70475581
Saskatchewan	1.64695192	0.88490102

Taking the MSE as an example, Quebec has the highest MSE, reaching 22.1956211, which indicates that the prediction error in Quebec is the largest. Conversely, British Columbia has the smallest MSE, only 1.138728416, indicating that the prediction error in British Columbia is the smallest.Similarly, we can also see that, taking MAE as an example, Quebec has the highest MAE, which is 2.704755812, indicating that the average prediction error in Quebec is the largest; whereas British Columbia has the smallest MAE, which is 0.753031308, indicating that the average prediction error in British Columbia is the smallest.

These results suggest that the predictive model performs worst in Quebec, which may be due to the data characteristics of Quebec, the adaptability of the model, or the existence of some outliers affecting the prediction results. On the contrary, the model performs best in British Columbia, which may be because the model is well adapted to the data characteristics of British Columbia.



Although we attempted to use deep learning models for predictions, our data volume is relatively small and does not meet the deep learning model's demand for many samples. Hence, we had to resort to using traditional machine learning methods such as RF and k-NN.

From the results, the predictive performance of RF and k-NN varies across cities. Overall, RF performs slightly better than k-NN in most cities. However, in certain specific cities, such as Saskatchewan and Manitoba, k-NN outperforms RF.

Additionally, in the RF model, the importance of COVID-19 cases and deaths varies from city to city. For instance, in Alberta, the number of cases has a larger impact on predicting the unemployment rate than the number of deaths. This indicates that we need to consider the specific problem when deciding which features to use. Although we could not utilize deep learning models for prediction, this study provides preliminary insights into how COVID-19 case and death statistics impact unemployment rates by comparing RF and k-NN models. Our findings assist in selecting more appropriate models for similar problems in the future and offer directions to improve predictive performance.

6. Conclusions & Problems Met

In the process of finishing this project, we did many tests, gave many hypotheses, and drew many conclusions. This is a list of the findings we had in the process:

1. The unemployment rate changed significantly when the pandemic was at its height. Ontario was the only province/territory where there is no significant difference in unemployment rate before the pandemic and now. (section 3.5)
2. The unemployment rate of both sexes reacted roughly the same to the pandemic. (section 3.3)
3. Different provinces/territories can react very differently to the pandemic, and it is therefore necessary to study the COVID-19 and unemployment data from different regions separately. (section 3.2)
4. The COVID-19 cases and deaths of a region is not linearly related to the region's unemployment rate, or very limitedly linearly related. (section 4.1)
5. Unemployment rate follows a different pattern to the unemployment rate when the pandemic was at its height. Limiting our time frame helps us capture that pattern. (section 4.3)
6. A region's unemployment rate is mainly affected by not only the COVID-19 cases and deaths of that region, but also by the overall pandemic situation of the entire country. (section 4.4)
7. The data of a region's COVID-19 cases and deaths, along with the entire country's COVID-19 cases and deaths, can explain 20% to 60% of the unemployment rate of that region. Newfoundland and Labrador is least directly affected by the COVID-19 cases and deaths, whereas Quebec is most directly affected. (section 4.5)
8. For a lesser amount of data, conventional statistical tests and regressions can give more accurate and more interpretable results compared to machine learning techniques. (section 5)

Apart from the findings, there are also things we can improve about our project, but didn’ t because of the scope and time budget of our data. Here is a list of possible improvements:

1. Canada has international trades with many countries. If we are able to consider the pandemic’s effect on Canada’s trading partners, we may be able to make more accurate models.
2. Different industries are affected by the pandemic in a different way. If we are able to subdivide the unemployment rate into the unemployment rates of each industry, we can possibly draw more conclusions.
3. A province/territory is more likely to trade with provinces/territories near it. If we can take into account a province/territory’s neighbors’ pandemic situation, maybe we can have more accurate results.
4. The loss of labor force or job opportunities due to the pandemic is less because of sick people being unable to offer or accept jobs, and more because of the regulations and quarantines in order to contain the pandemic. If we can evaluate the policies of a region in response to the pandemic along with the COVID-19 cases and deaths, we can possibly give more insightful results.
5. We can often see that the unemployment rate does not go down as soon as the COVID-19 situation gets better. This is possibly because the economy needs time to recover. Hence, the unemployment of a month may be the result of pandemic figures of previous months. If we are able to analyze the time series as a whole, instead of as separate (COVID-19 data, unemployment data) pairs of each month, we may be more successful.

7. Project Experience Summary

Eric：Cleaned the Canada labor force data to an unemployment rate-only CSV file. Processing the data, visualizing the data, and testing several hypotheses.

Isaac: Proposed the project idea. Combined the cleaned data for the unemployment and COVID data into a single csv file. Performed different linear, multilinear and polynomial regressions on the combined unemployment and pandemic data, and analyzed the results.

Kaize Gu：Cleaned the data, using machine learning methods, specifically Random Forests (RF) and k-Nearest Neighbors (k-NN), to make predictions on given data.(use COVID data predicate unemployment rate)

REFERENCES PAGE

Government of Canada, Statistics Canada. (2023, July 7). *Add/Remove data - Labour force characteristics by province, monthly, seasonally adjusted*. https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1410028703

Public Health Infobase - Data on COVID-19 in Canada - Open Government portal. (n.d.). https://open.canada.ca/data/en/dataset/261c32ab-4cfd-4f81-9dea-7b64065690dc

WHO Coronavirus (COVID-19) Dashboard. (n.d.). WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. https://covid19.who.int/