Final Project Report

CMPT 353 Computational Data Analysis, Summer 2023

Isaac Ding 301425524

Huanyu Zhou 301417467

Kaize Gu 301416566

1. Introduction

1.1 Summary of Project

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

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.

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? Does the unemployment rate differ in different COVID-19 pandemic periods?
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? Are the covid cases/deaths linearly related to the unemployment rate? How much do 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?

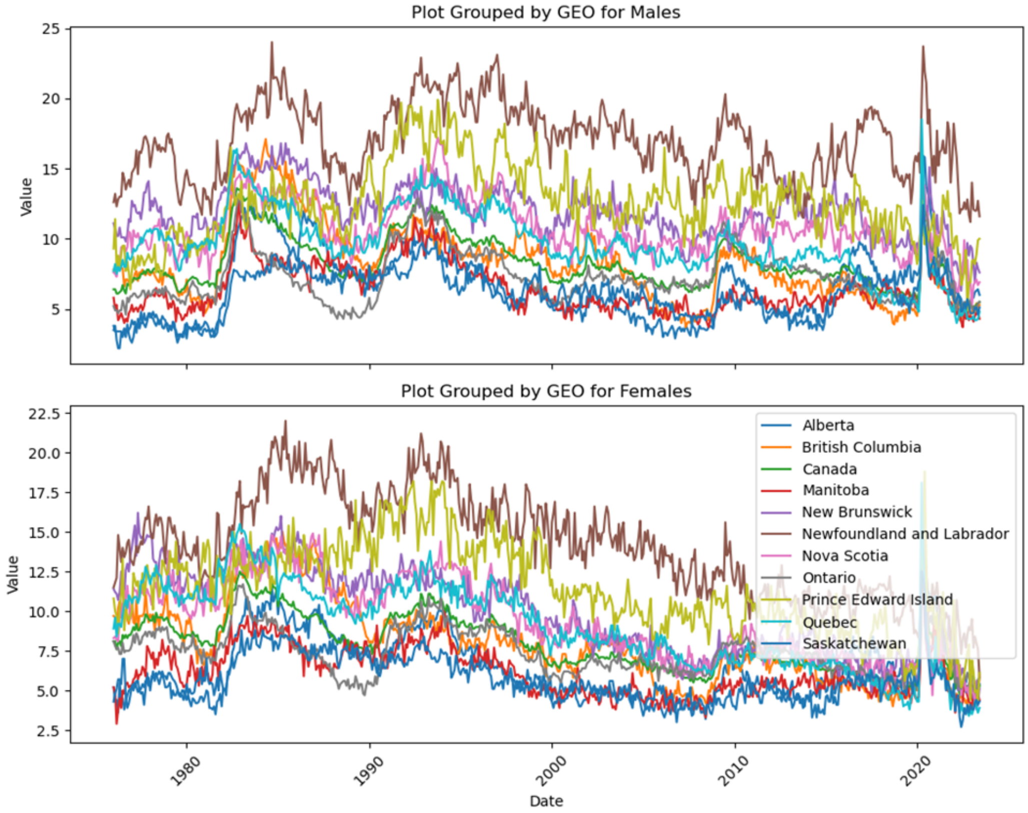
2. Data Gathering & Cleaning

In this project, we explore 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 original data are a combination of different meanings of data such as “Population”, “Full-time employment rate” etc. After filtering the data to value mean= ‘Unemployment rate’

only, 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.

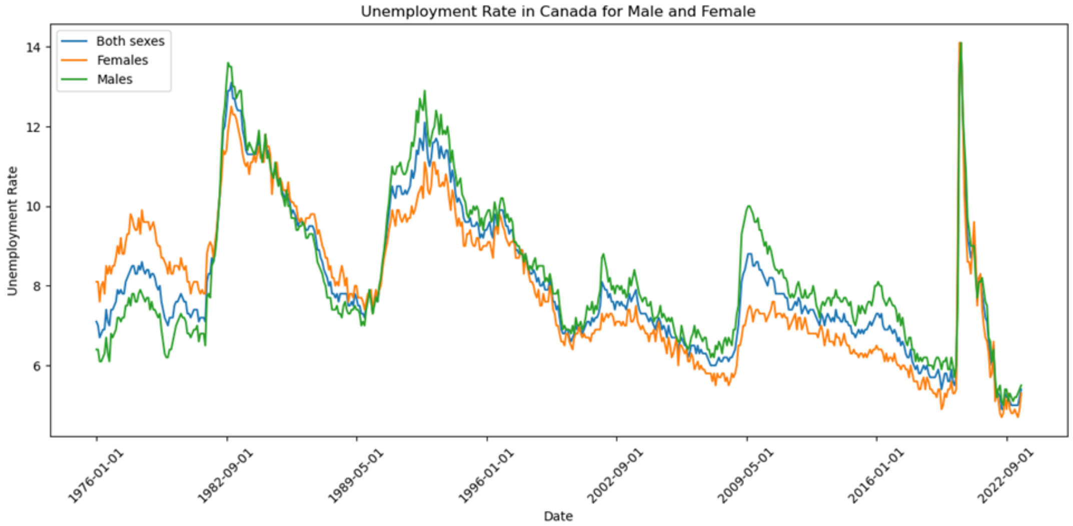
3. Data Processing & Statistical Tests

3.1 Visualizing Unemployment Rate from 1976-2023



After visualizing the data, 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.

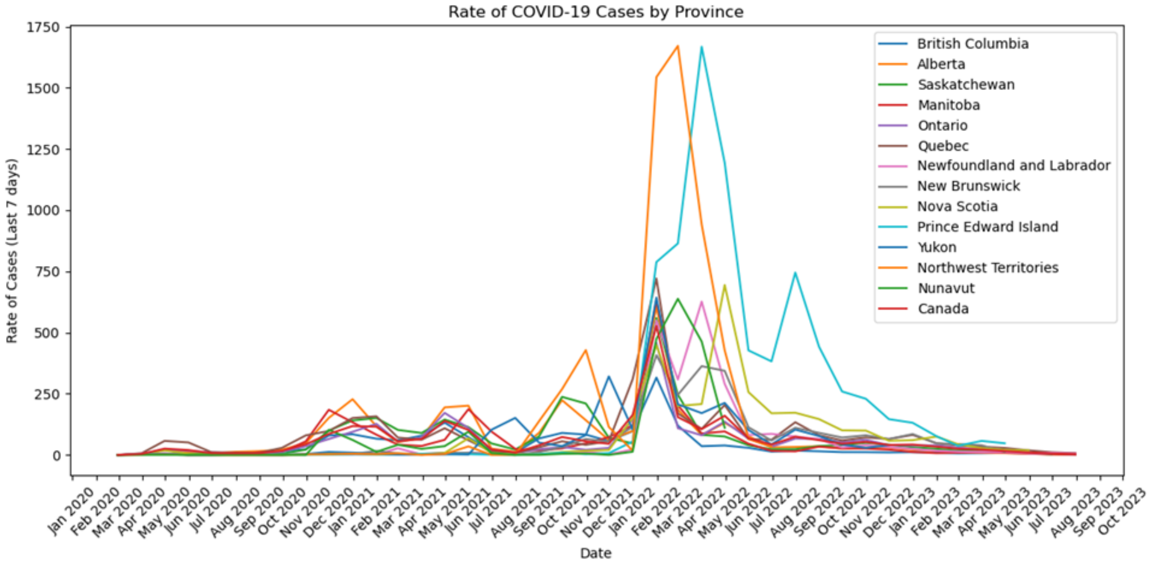
3.2 Visualizing Unemployment Rate for Different Sexes



In addition to males and females, we also visualized the plot for Both sexes. As the male and female unemployment rate fluctuates, they follow very similar trends, and are very close to the unemployment rates for both sexes. We think it is reasonable to represent male and female unemployment rates 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.

3.3 Visualizing Covid-19 Weekly Cases

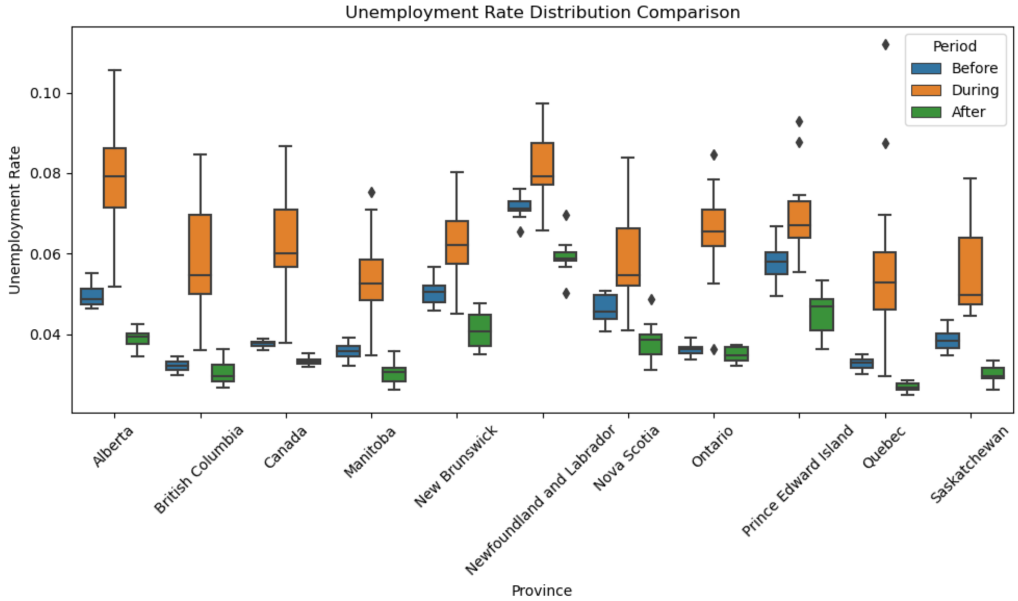


From the plot we can see the case rate for all the provinces significantly increased in Dec 2021 and reached a peak at around Feb 2022 to Apr 2022.

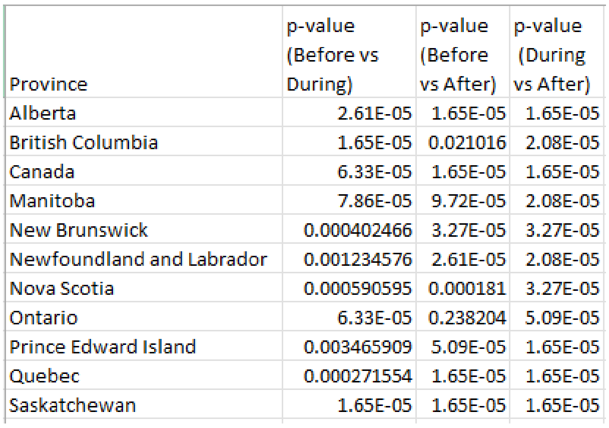
Alberta seems to have a relatively larger statistic than other provinces, so we tested if there's a significant difference between data in Alberta and a random province, Yukon. The result had p_value = 0.9955263219772348. Therefore the data in Alberta has no significant difference from the data in Yukon.

3.4 Unemployment rate comparison Before, During, and After the Pandemic

We chose 2019-01 – 2020–01 as data before the pandemic, 2020-02 – 2021-02 as data during the pandemic, and 2022-06 – 2023-07 as data after the pandemic.



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:



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

Except for 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. Therefore in all the regions in Canada, 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

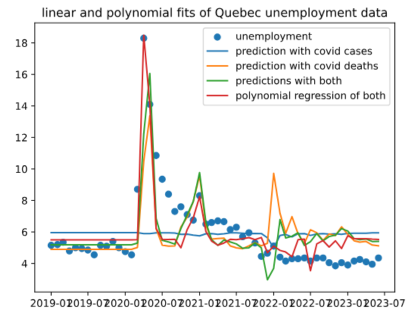
4.1 Linear Correlation Test

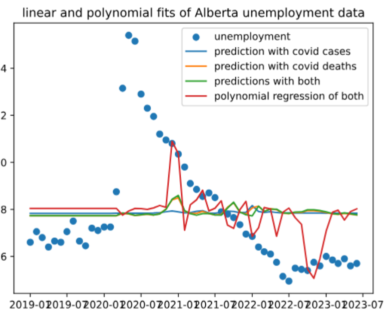
We established that the COVID-19 pandemic does have an effect on the unemployment rate across Canada. The next thing is to consider if a linear relationship exists. We performed Pearson's correlation tests for the scope of the entire Canada. The unemployment rate has Pearson’s correlation value of about -0.08 with the number of covid cases and about 0.05 with the number of covid deaths. This means there is a very weak linear relation between these two covid data and the unemployment rate.

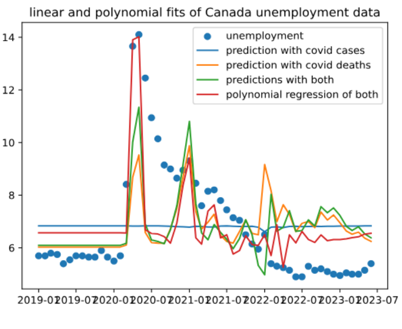
4.2 Trial 1: Fitting a Region’s COVID-19 Data and Unemployment Rate

The covid data and unemployment aren’t linearly related to the unemployment rate, so simply performing linear regression may give inaccurate results. Therefore we also tried multilinear and polynomial regression, which provide more flexibility to the model.

performing linear regression with input as 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 multilinear and polynomial regression, sklearn models and therefore we measured the training score instead of the p-value. We used both covid cases and deaths as input. In general, all models didn’t perform very well. Here is a visualization of the model’s predictions

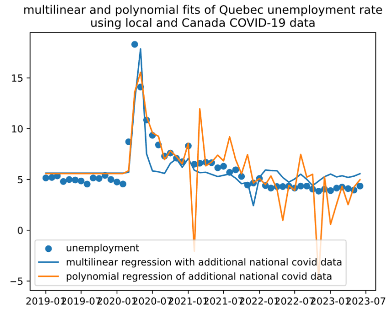


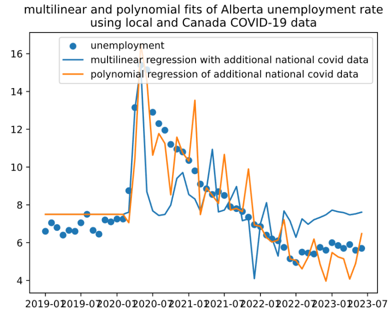


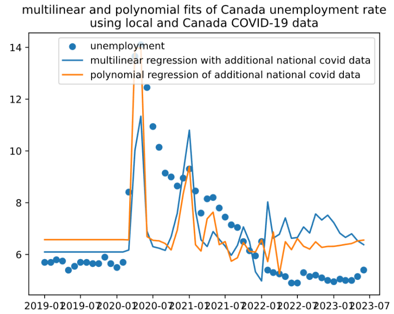


4.3 Trail 2: improving the performance of the models.

We tried two ways to improve the performance of the model: we tried limiting the data points between 2020-01 and 2022-12, and we also tried adding the countrywide covid data as extra input features when fitting the local covid data and unemployment. The linear regressions with only one input feature, either covid cases and covid deaths, had worse predictions. But the multilinear regressions and the polynomial regressions have improved in prediction and accuracy for both attempts. The best performance was with adding the countrywide covid data and using the multilinear regression, where we had 0.20 to 0.63 accuracy on different regions. Here's how they look in comparison:







Adding the national covid data gave the polynomial model 4 input features, and our data is probably not enough to train it, causing the unstable predictions.

The fact that limiting the timespan and adding the countrywide data improved accuracy showed that the unemployment rate follows a different trend during the pandemic, and also that the national pandemic also impacts the local unemployment rate.

4.4 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. Performing regression is actually finding coefficients that relate the input features to output features. If we are able to find coefficients that give our model a 90% accuracy, a linear combination or polynomial of our input features X is 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 after adding countrywide covid data as extra columns to predict local unemployment, 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, the region is less directly affected by the COVID-19 cases and deaths. This means, 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 prediction.

The features include 'covid cases' and 'covid deaths' and the target variable is 'total unemployment rate(%)'.

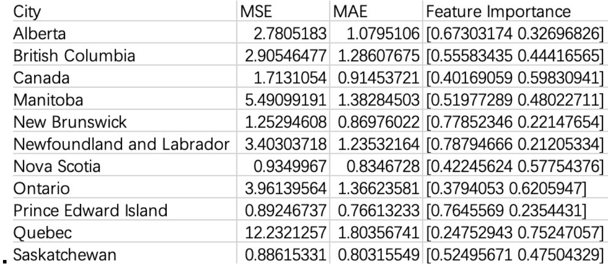
We used the two machine learning methods, Random Forest and k-Nearest Neighbors:

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

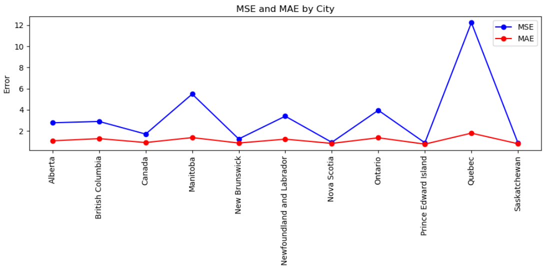
k-Nearest Neighbors (k-NN): k-NN is an instance-based learning or a local approximation and simplification method. We set k = 5 and standardized the features.

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)

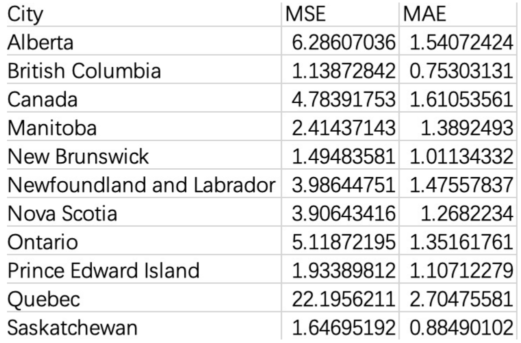


City	MSE	MAE	Feature Importance
Alberta	2.7805183	1.0795106	[0.67303174 0.32696826]
British Columbia	2.90546477	1.28607675	[0.55583435 0.44416565]
Canada	1.7131054	0.91453721	[0.40169059 0.59830941]
Manitoba	5.49099191	1.38284503	[0.51977289 0.48022711]
New Brunswick	1.25294608	0.86976022	[0.77852346 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]

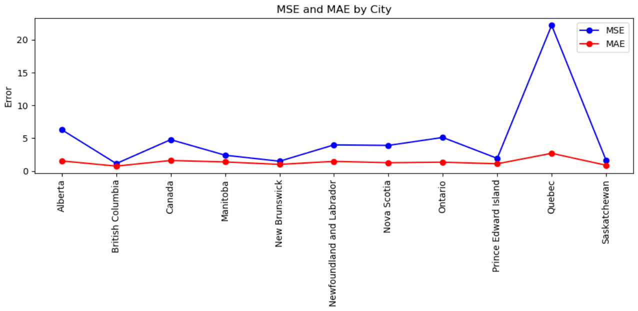


For instance, Alberta has an MSE of 2.780518299 and an MAE of 1.079510596. The importance of 'covid cases' and 'covid deaths' in the prediction model for Alberta are 0.67303174 and 0.32696826 respectively, suggesting that the number of COVID-19 cases is still more influential than the number of deaths when predicting the unemployment rate in this city.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.

The results showed that RF generally performed better than k-NN, though k-NN outperformed RF in specific cities like Saskatchewan and Manitoba.Although deep learning wasn't suitable since the data was too small, our comparison of RF and k-NN provides insights into how COVID-19 statistics impact unemployment rates and assists in choosing better models for similar problems in the future.

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 during the pandemic. Ontario was the only province/territory where there is no significant difference in the unemployment rate before and after the pandemic.(section 3.5)
2. The unemployment rate of both sexes reacted roughly the same to the pandemic. (section 3.2)
3. The COVID-19 cases and deaths in a region are not linearly related to the region's unemployment rate, or very limitedly linearly related. (section 4.1)
4. The unemployment rate follows a different trend during the pandemic, and also that the national pandemic also impacts the local unemployment rate. (section 4.3)
5. The national and regional covid data can explain 20% to 60% of the unemployment rate of a region. Newfoundland and Labrador is least directly affected by the COVID-19 cases and deaths, whereas Quebec is most directly affected. (section 4.5)
6. For a lesser amount of data, 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 trade with many countries. If we are able to consider the pandemic's effect on Canada's trading partners, we can make more accurate models.
2. Different industries are affected by the pandemic in different ways. If we are able to subdivide the unemployment rate into the unemployment rates of each industry, we can possibly draw more conclusions.
3. The unemployment rate does not go down or up as soon as the COVID-19 situation gets better or worse. This is possibly because the economy needs time to recover. Hence, the unemployment rate of a month may be the result of the pandemic figures of previous months. If we are able to analyze data in relation to the data before it, we may be more successful.

REFERENCES

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/

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)