Hypothesis:

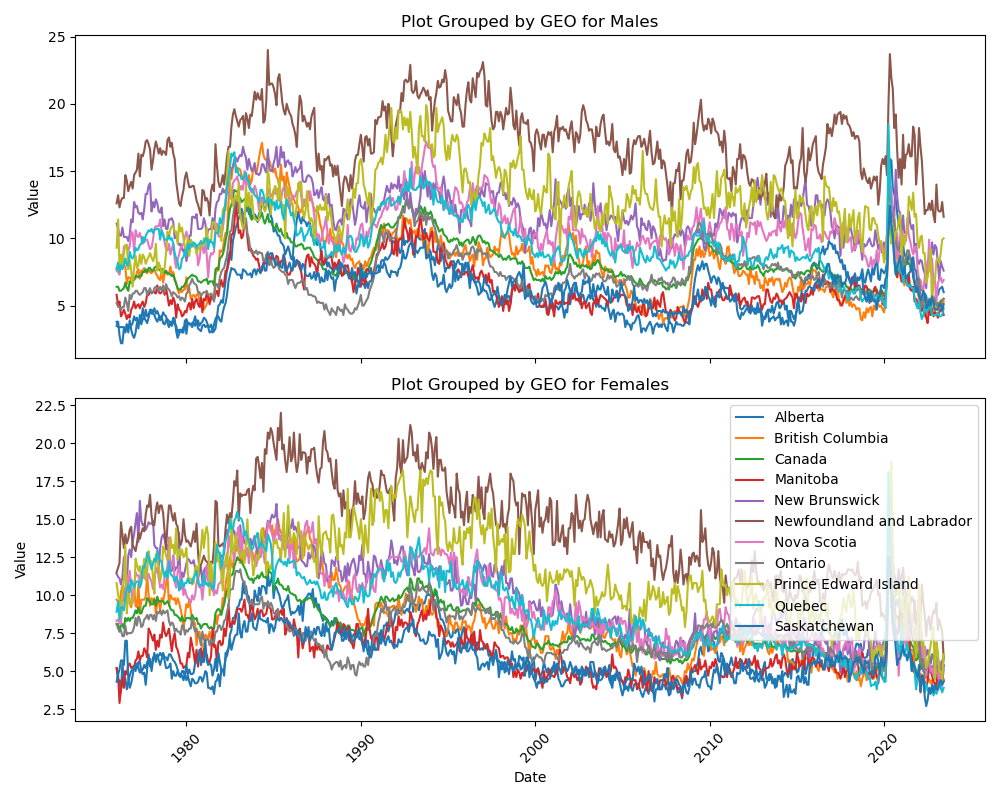
* Is the unemployment rate different for male and female?
* Did covid-19 case rate affect unemployment rate?
* Can we use covid-19 case rate to predict unemployment rate?

Data clean:

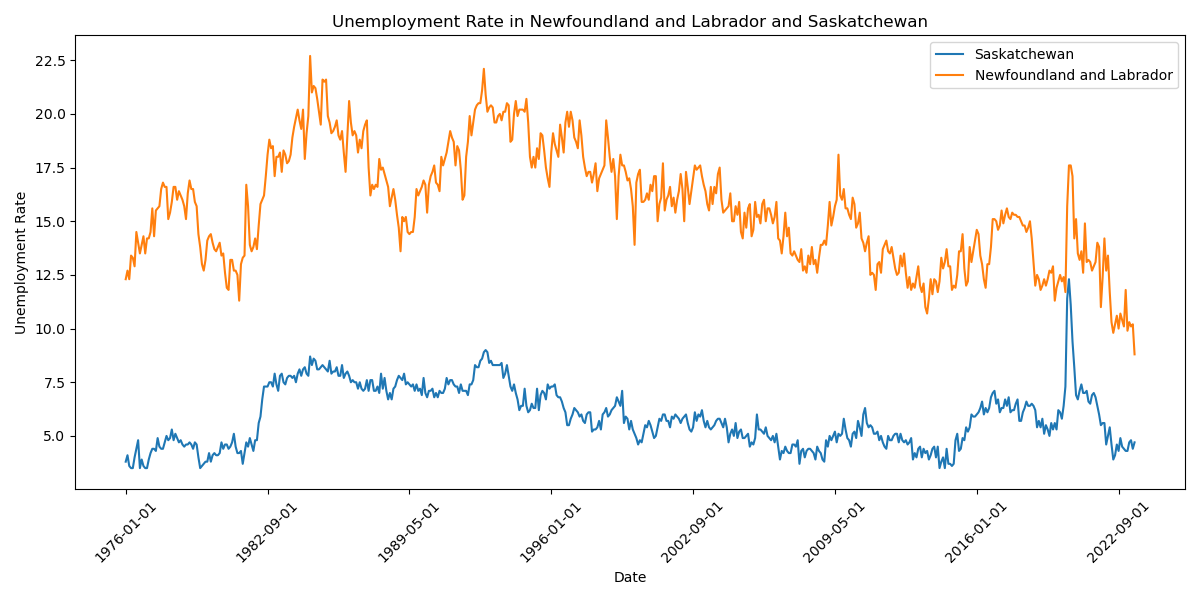
In this project, we are focusing on the relationship of unemployment rate data and covid-19 case data. However, the Labour force statistics provided by Statistics Canda is too chaotic to provide the data we need, so we did the following data cleaning:

The combined data are not separately performing the value of each parameter, it store the meaning of data to an attribute named ‘Labour force characteristics’. 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 ‘Clean&Visualization.py’ file.

Data visualization of Unemployment rate from 1976-2023

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

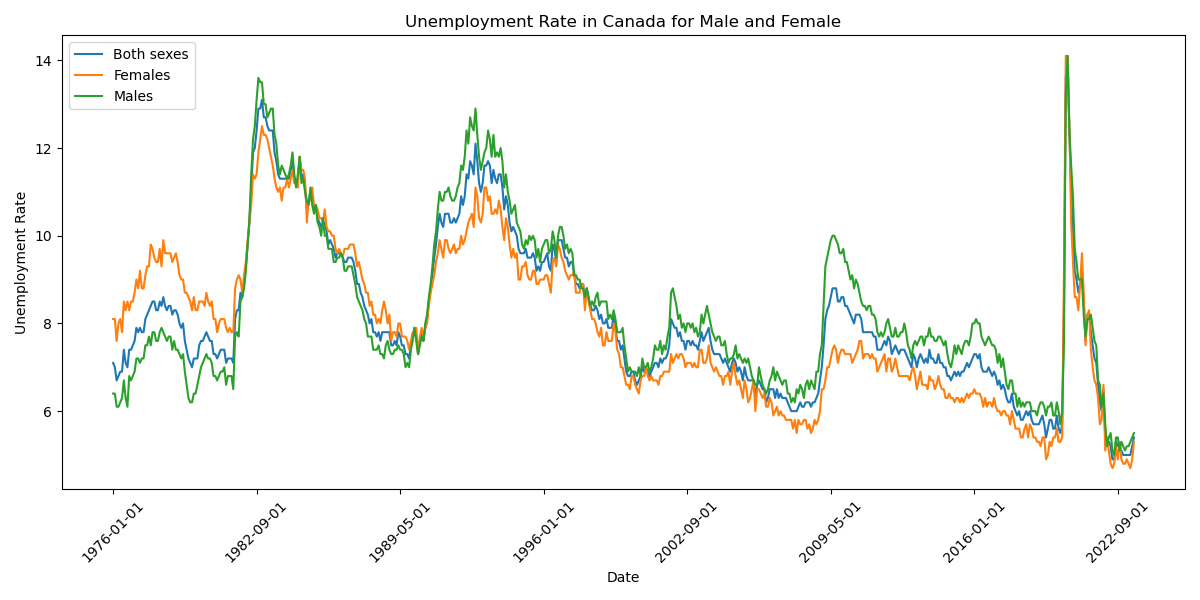
Data visualization for both sexes in 'Newfoundland and Labrador', and 'Saskatchewan' from 1976 – 2023

As shown in plot, the unemployment rate in 'Newfoundland and Labrador' is always higher than that in 'Saskatchewan', however, to further confirm our hypothesis, we applied mannwhitneyu test on it and get the output p\_value = 1.351418785610596e-187.

As the p\_value < 0.05, we can reject null hypothesis, and conclude there is significant difference between the data in these two regions.

The code is provided in ‘GEO\_comparision.py’ file.

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

Data visualization for both sex in Canda from 1976 – 2023 for different sex:

Except males and females, we also visualized the plot for ‘Both sexes’, however, as the data fluctuates, we can not draw a conclusion from this plot, we also provided a data summarization for both genders.

Unemployment rate summary of Male:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GEO | Mean | Median | Standard Deviation | Minimum | Maximum |
| Alberta | 6.654035 | 6.3 | 2.540254 | 2.9 | 15.9 |
| British Columbia | 8.282456 | 7.9 | 2.782207 | 3.7 | 17.1 |
| Canada | 8.294211 | 7.7 | 1.838186 | 5 | 14.1 |
| Manitoba | 6.49193 | 5.9 | 1.747246 | 3.7 | 12.8 |
| New Brunswick | 11.96614 | 11.9 | 1.869549 | 6.3 | 16.8 |
| Newfoundland and Labrador | 16.98333 | 17.05 | 2.614597 | 11.2 | 24 |
| Nova Scotia | 10.82158 | 10.4 | 1.917932 | 5.2 | 17.2 |
| Ontario | 7.42386 | 7 | 1.900539 | 4.3 | 13.7 |
| Prince Edward Island | 12.56561 | 12.3 | 2.628486 | 4.9 | 19.9 |
| Quebec | 9.68807 | 9.4 | 2.385361 | 4 | 18.5 |
| Saskatchewan | 6.072105 | 6.2 | 1.60335 | 2.2 | 12.4 |

Unemployment rate summary of Female:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GEO | Mean | Median | Standard Deviation | Minimum | Maximum |
| Alberta | 6.439474 | 5.9 | 2.060287 | 3 | 14.9 |
| British Columbia | 8.038246 | 7.7 | 2.605546 | 3.5 | 15 |
| Canada | 7.932456 | 7.6 | 1.78325 | 4.7 | 14.1 |
| Manitoba | 6.128772 | 5.8 | 1.519798 | 2.9 | 11.8 |
| New Brunswick | 9.875088 | 9.6 | 2.707364 | 3.8 | 16.2 |
| Newfoundland and Labrador | 14.19649 | 14.2 | 3.232136 | 5.9 | 22 |
| Nova Scotia | 9.432807 | 9.15 | 2.538171 | 4.5 | 16.1 |
| Ontario | 7.444035 | 7.3 | 1.500817 | 4.7 | 14.4 |
| Prince Edward Island | 11.50579 | 11.2 | 2.789445 | 4.4 | 18.8 |
| Quebec | 9.064386 | 9.05 | 2.717666 | 3.4 | 18.1 |
| Saskatchewan | 5.732281 | 5.5 | 1.372019 | 2.7 | 12.5 |

From the summary, we can see the mean unemployment rate for male is relatively higher than that data for female, but we still need to further confirm our hypothesis, so we introduced t test.

This time, I just kept the data for the region of Canada, and compare it for different genders, the following is the p\_value of the test:

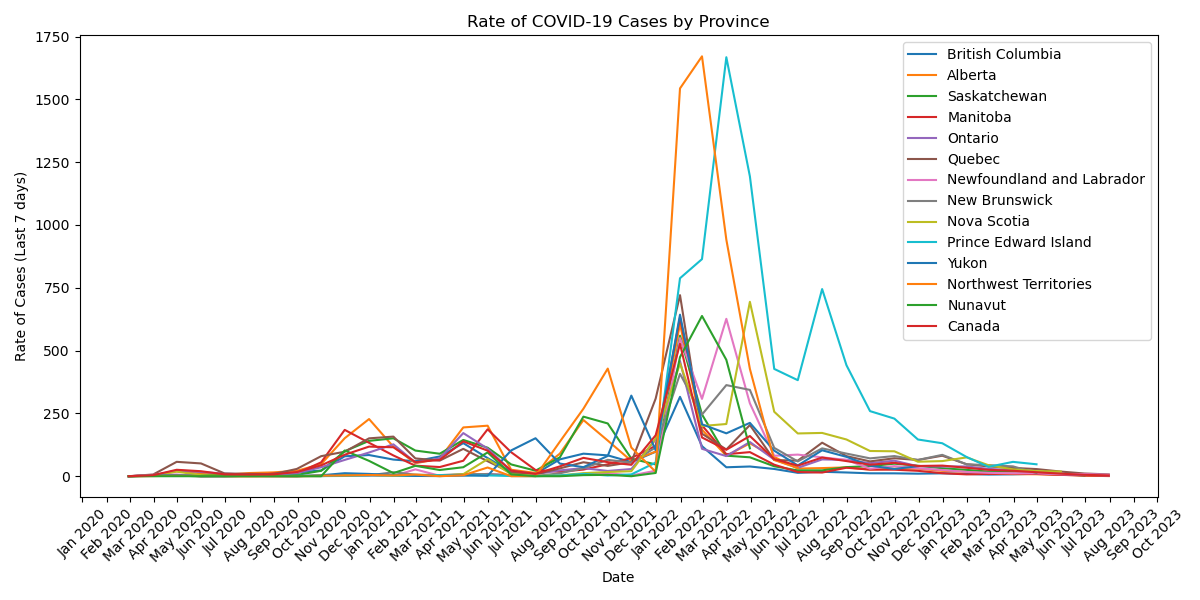
p\_value = 0.0007703222294190954

As p\_value < 0.05, we can conclude there is a significant difference in unemployment rates between male and female.

The code is provided in ‘Unempl\_Sex\_Analysis.py’ file.

Before we start analysis the relationship between the selected two dataset, we first analysis the covid-19 average case rate.

Data visualization of Covid-19 average case rate (last 7) from Feb 2020-July 2023



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

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

Output:

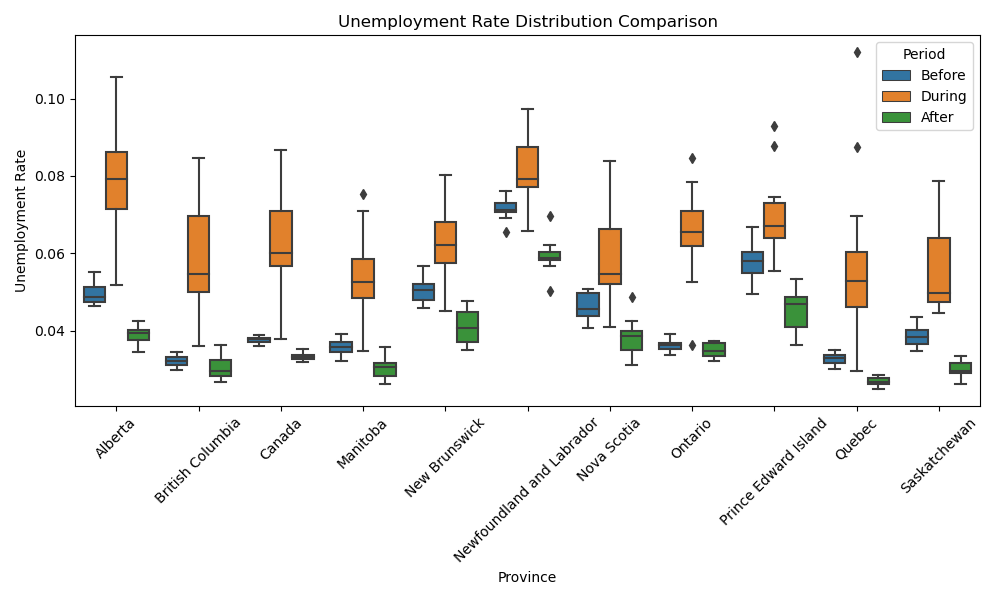
p\_value = 0.9955263219772348

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 period, we split our unemployment data to three subset, Jan 2019 – Jan 2020 represent data before pandemic, Feb 2020 – Feb 2021 represent data during pandemic, June 2022 – June 2023 represent data after pandemic.

We first create a box plot for data visualization.

Unemployment rate comparison among ‘Data before pandemic’, ‘Data during pandemic’, ‘Data after pandemic’



As shown in plot, the unemployment rate data during pandemic is much higher than that before pandemic and after pandemic in all the provinces in Canada. To further confirm our hypothesis, we did the mannwhitneyu test among each indicators 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 | 0.021016256 | 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 | 0.00018141 | 3.27E-05 |
| Ontario | 6.33E-05 | 0.238204239 | 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 |