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:

First, the data was stored in five folders represented different years from 2019 – 2023 and in each folders there are 12 .csv files that represented different months in each year. We wrote a python file named combine.py to read the folders, extract all the data and output a combined csv which included all the data since 2019.

Second, 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’ only, we are able to do the data for unemployment rate trend and the visualization and hypothesis test for is there different between unemployment rate for male and female.

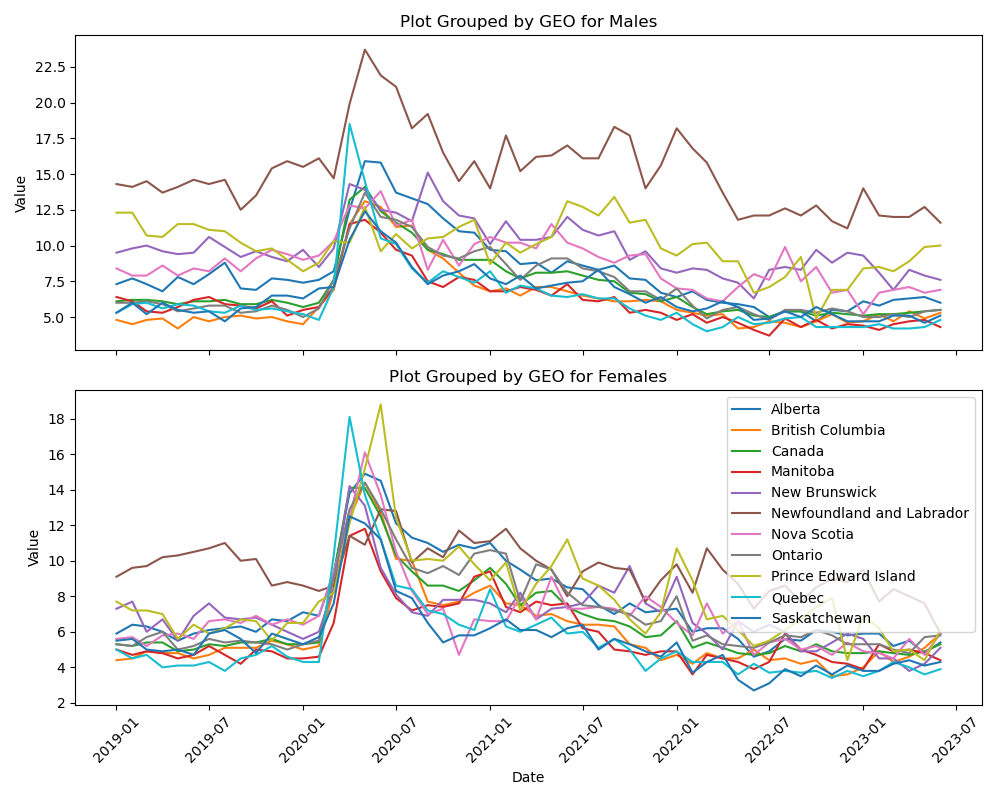
Third, when focusing on the relationship between unemployment rate and covid-19, we don’t care about gender too much, so we need to combine the data for different genders for further use. However, combine the gender data are much harder than get the average unemployment rate for males and females because of the different population. So we group the data by ‘GEO’ and ‘REF\_DATE’, calculate the sum of population and unemployment population. Use the number of people who are unemployed divided by the total population, we can get the combined data we want. The whole procedure was provided in Combine\_male\_female.py.

Unemployment rate data summary for female

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GEO | Mean Unemployment | Median Unemployment | Standard Deviation | Minimum Unemployment | Maximum Unemployment |
| Alberta | 8.107407407 | 7.6 | 2.62584052 | 4.7 | 15.9 |
| British Columbia | 6.209259259 | 5.25 | 2.220425636 | 4.2 | 13.1 |
| Canada | 7.02962963 | 6.1 | 2.178334973 | 5 | 14.1 |
| Manitoba | 6.096296296 | 5.7 | 1.80763193 | 3.7 | 11.8 |
| New Brunswick | 9.861111111 | 9.6 | 1.855986405 | 6.3 | 15.1 |
| Newfoundland and Labrador | 15.15185185 | 14.6 | 2.744582364 | 11.2 | 23.7 |
| Nova Scotia | 8.812962963 | 8.6 | 1.80031929 | 5.2 | 13.8 |
| Ontario | 7.022222222 | 5.85 | 2.208103999 | 4.8 | 13.7 |
| Prince Edward Island | 9.974074074 | 10.1 | 1.743579617 | 4.9 | 13.4 |
| Quebec | 6.211111111 | 5.55 | 2.530468423 | 4 | 18.5 |
| Saskatchewan | 6.52962963 | 6 | 1.711407854 | 4.6 | 12.4 |

Unemployment rate data summary for male

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GEO | Mean Unemployment | Median Unemployment | Standard Deviation | Minimum Unemployment | Maximum Unemployment |
| Alberta | 8.107407 | 7.6 | 2.625841 | 4.7 | 15.9 |
| British Columbia | 6.209259 | 5.25 | 2.220426 | 4.2 | 13.1 |
| Canada | 7.02963 | 6.1 | 2.178335 | 5 | 14.1 |
| Manitoba | 6.096296 | 5.7 | 1.807632 | 3.7 | 11.8 |
| New Brunswick | 9.861111 | 9.6 | 1.855986 | 6.3 | 15.1 |
| Newfoundland and Labrador | 15.15185 | 14.6 | 2.744582 | 11.2 | 23.7 |
| Nova Scotia | 8.812963 | 8.6 | 1.800319 | 5.2 | 13.8 |
| Ontario | 7.022222 | 5.85 | 2.208104 | 4.8 | 13.7 |
| Prince Edward Island | 9.974074 | 10.1 | 1.74358 | 4.9 | 13.4 |
| Quebec | 6.211111 | 5.55 | 2.530468 | 4 | 18.5 |
| Saskatchewan | 6.52963 | 6 | 1.711408 | 4.6 | 12.4 |

Data visualization of Unemployment rate from 2019-2023

As shown in the plot, the unemployment rate experienced an significant increase at January 2020 for both male and female in all the provinces in Canada and reached a peak at around May 2020.

Hypothesis Testing about if there’s a difference between unemployment rate of Male and Female in the region of Canada.

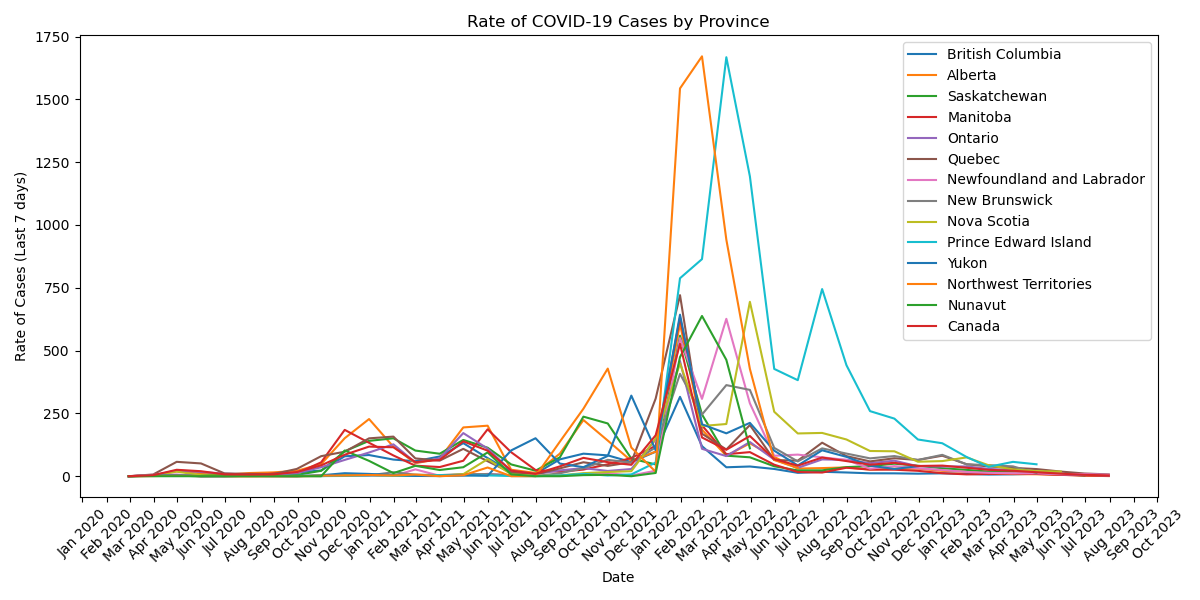
Output:

p\_value = 0.3335291231595887

There is no significant difference in unemployment rates between male and female.

As there is no significant difference between data for male and female, we are safe to combine these two datasets for future analysis.

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.

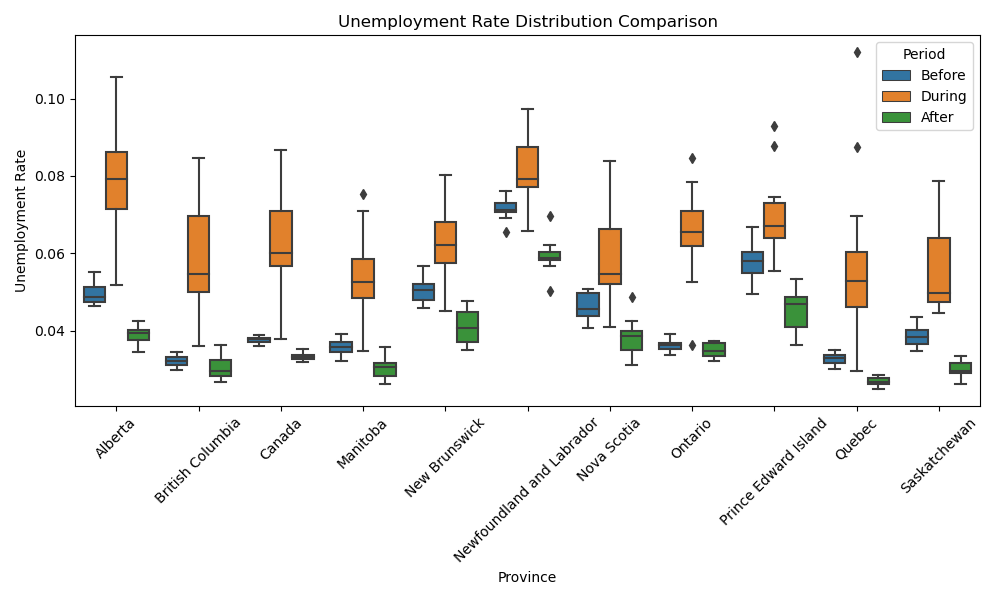
As the data from Alberta seems has a relatively larger statistic than other province, we did a hypothesis test to test if there’s significant difference between data in Alberta and the other province we choose which is Yukon.

Output:

p\_value = 0.9955263219772348

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

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



|  |  |  |  |
| --- | --- | --- | --- |
| 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 |