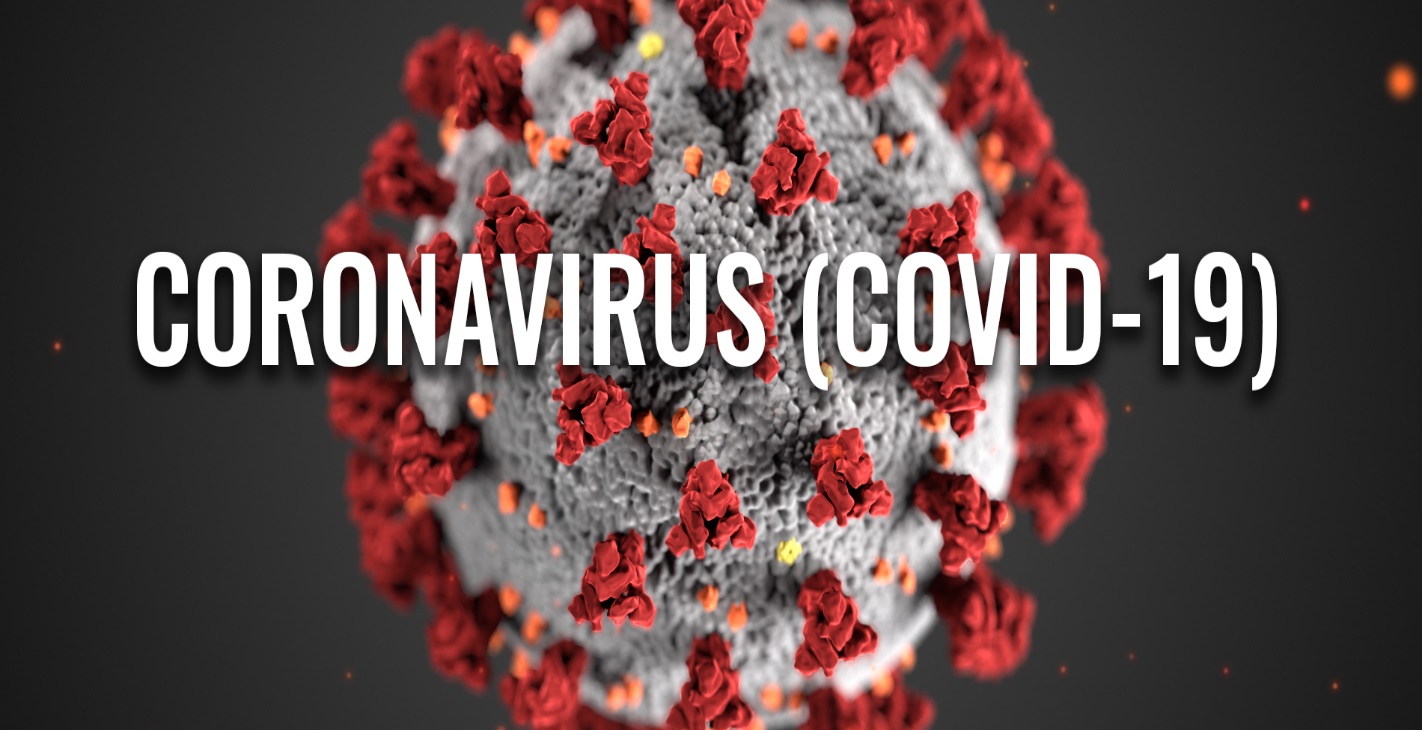
**COVID-19 Pandemic Report**

Problem Statement

Findings

Conclusions

Random Forest



**Submitted By: Submitted To:**

**Priyanshi Chakraborty Asem Omari**

**Jeena Helen Francis**

**Nibhashri Satyen**

**Manoj Moond**

**Parth Patel**

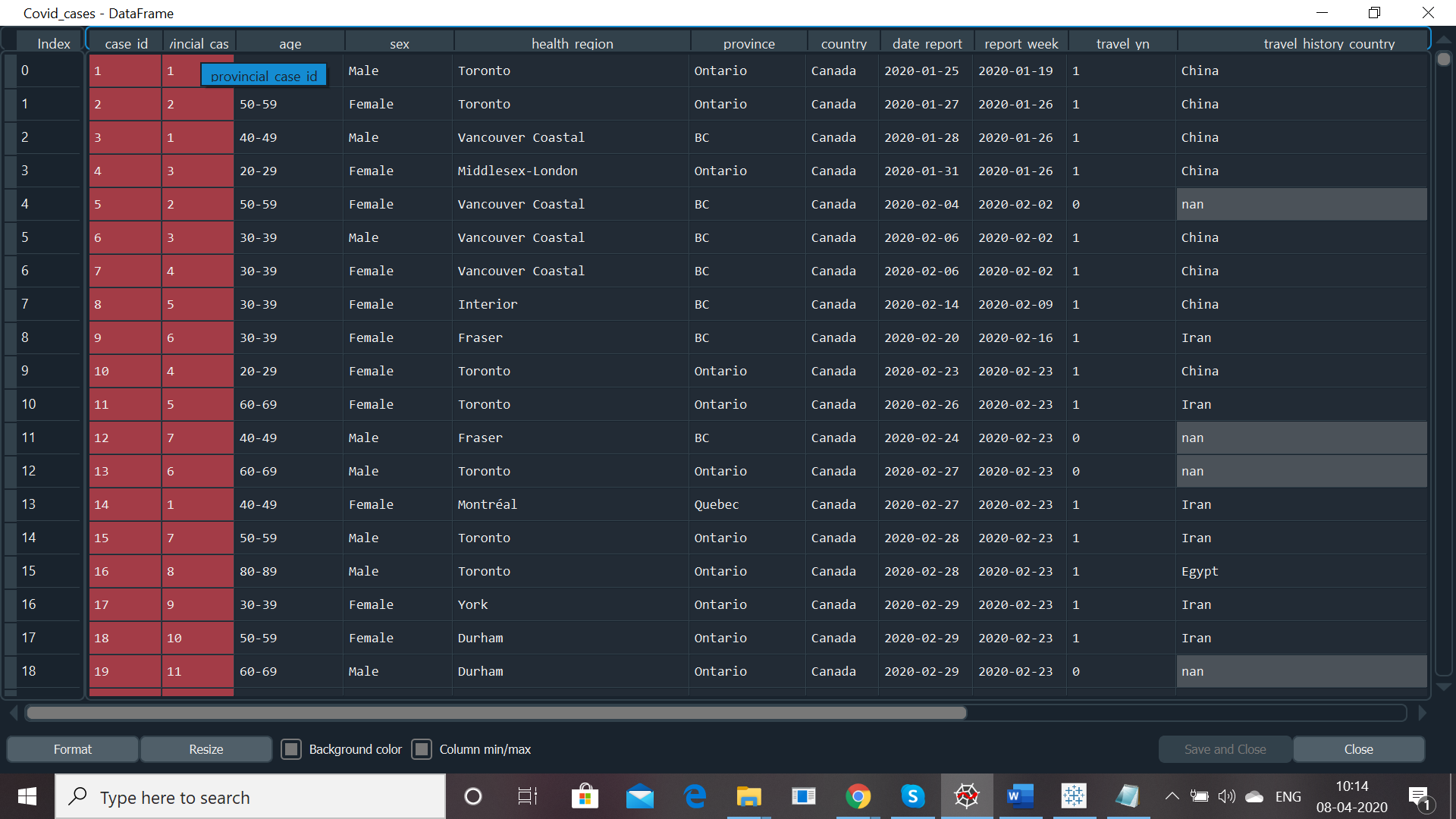
2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it’s unclear how easily or sustainably this virus is spreading between people – CDC.

**As of March 2020, below is the data for Canada we are using:**

**Cases: This file records the daily new cases.**

**Attributes:**

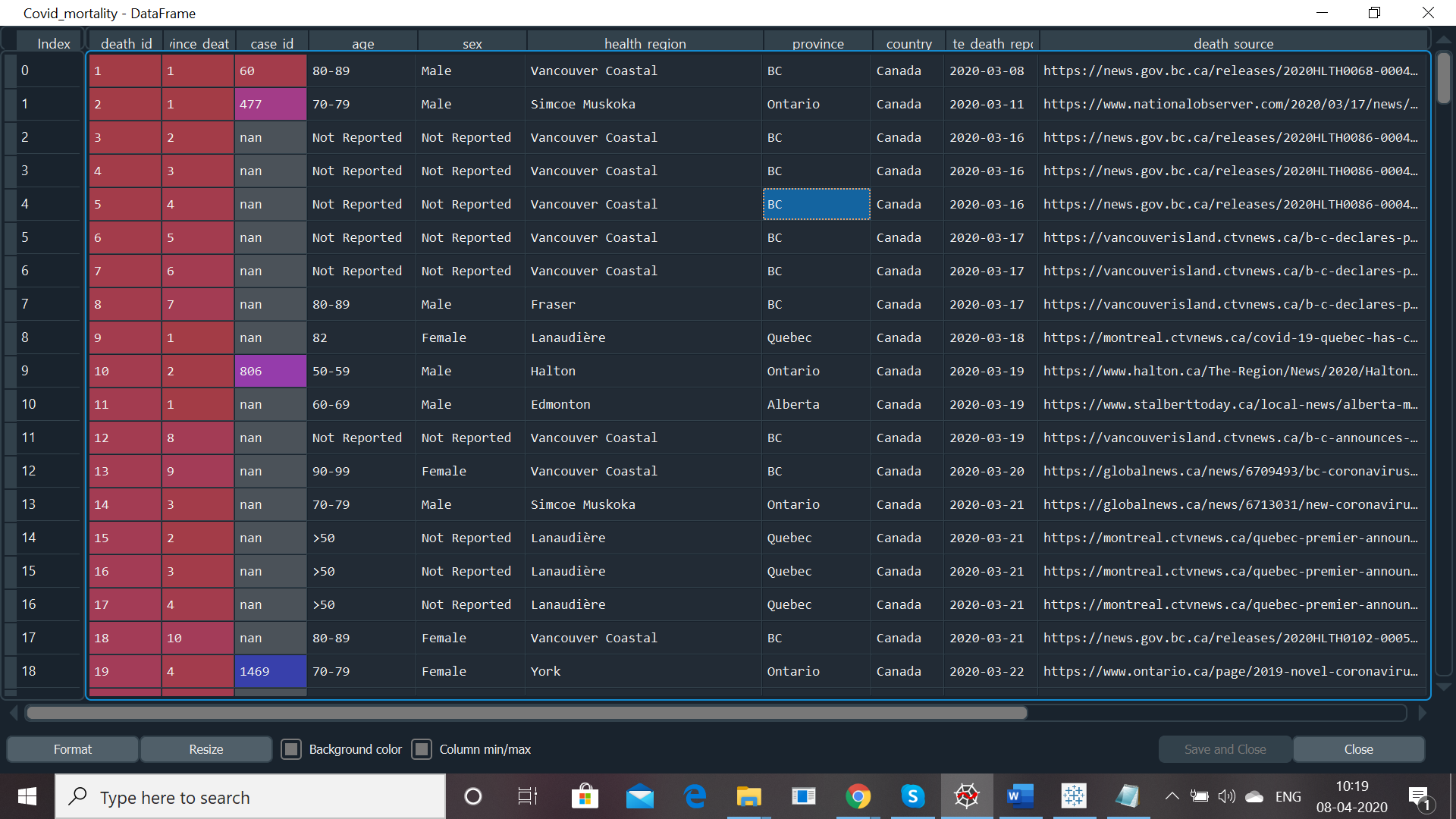
* case\_id
* provincial\_case\_id
* age
* sex
* health\_region
* province
* country
* date\_report
* report\_week
* travel\_yn
* travel\_history\_country
* locally\_acquired
* case\_source
* additional\_info
* additional\_source
* method\_note



**Mortality: This file records the death information**

**Attribute:**

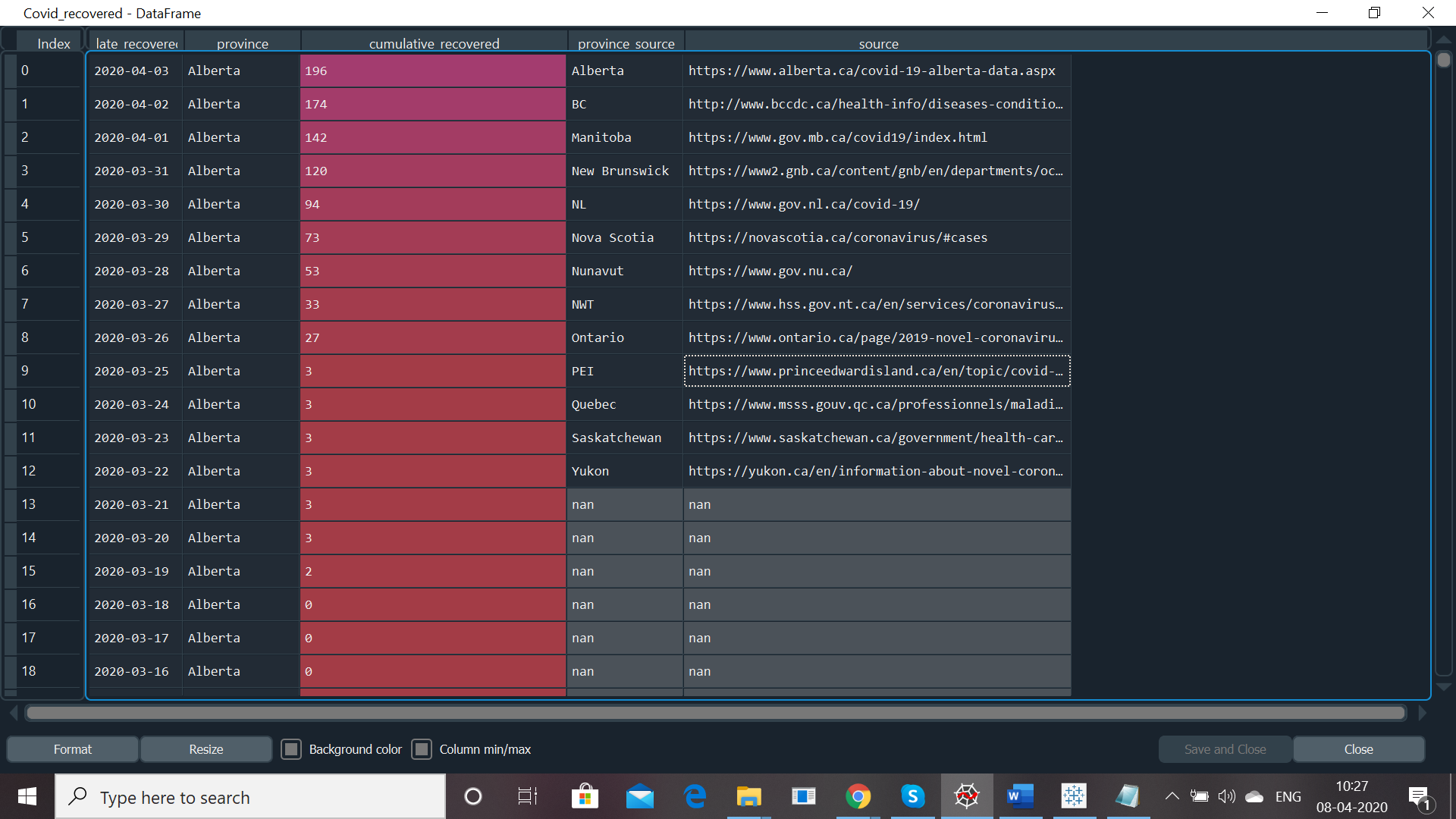
* death\_id
* province\_death\_id
* case\_id
* age
* sex
* health\_region
* province
* country
* date\_death\_report
* death\_source
* additional\_info
* additional\_source

r

**Recovered: This file records the recovered patients.**

**Attributes:**

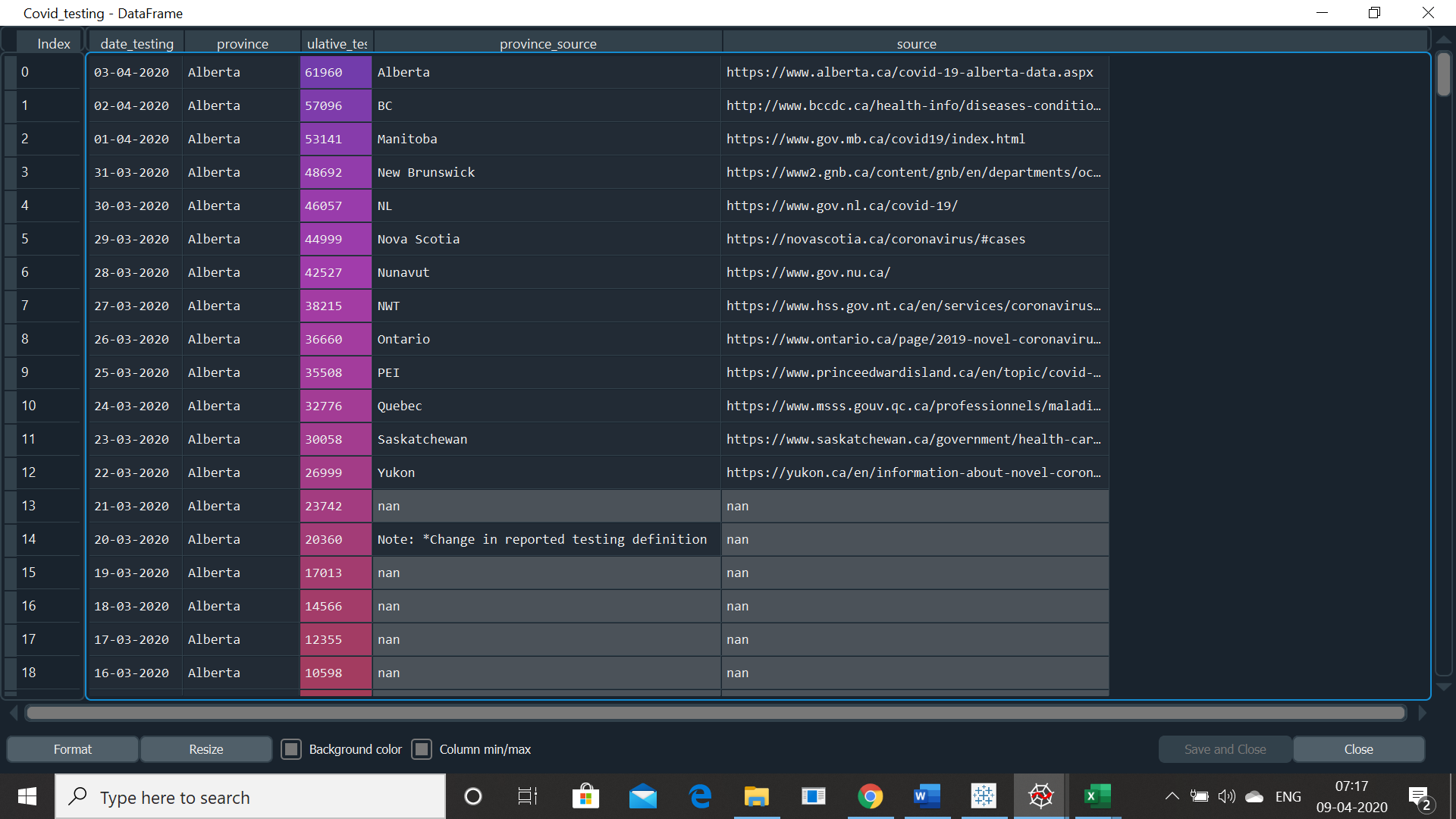
* date\_recovered
* province
* cumulative\_recovered
* province\_source
* source



**Testing: This file records the testing progress in each province.**

**Attributes:**

* date\_testing
* province
* cumulative\_testing
* province\_source
* source

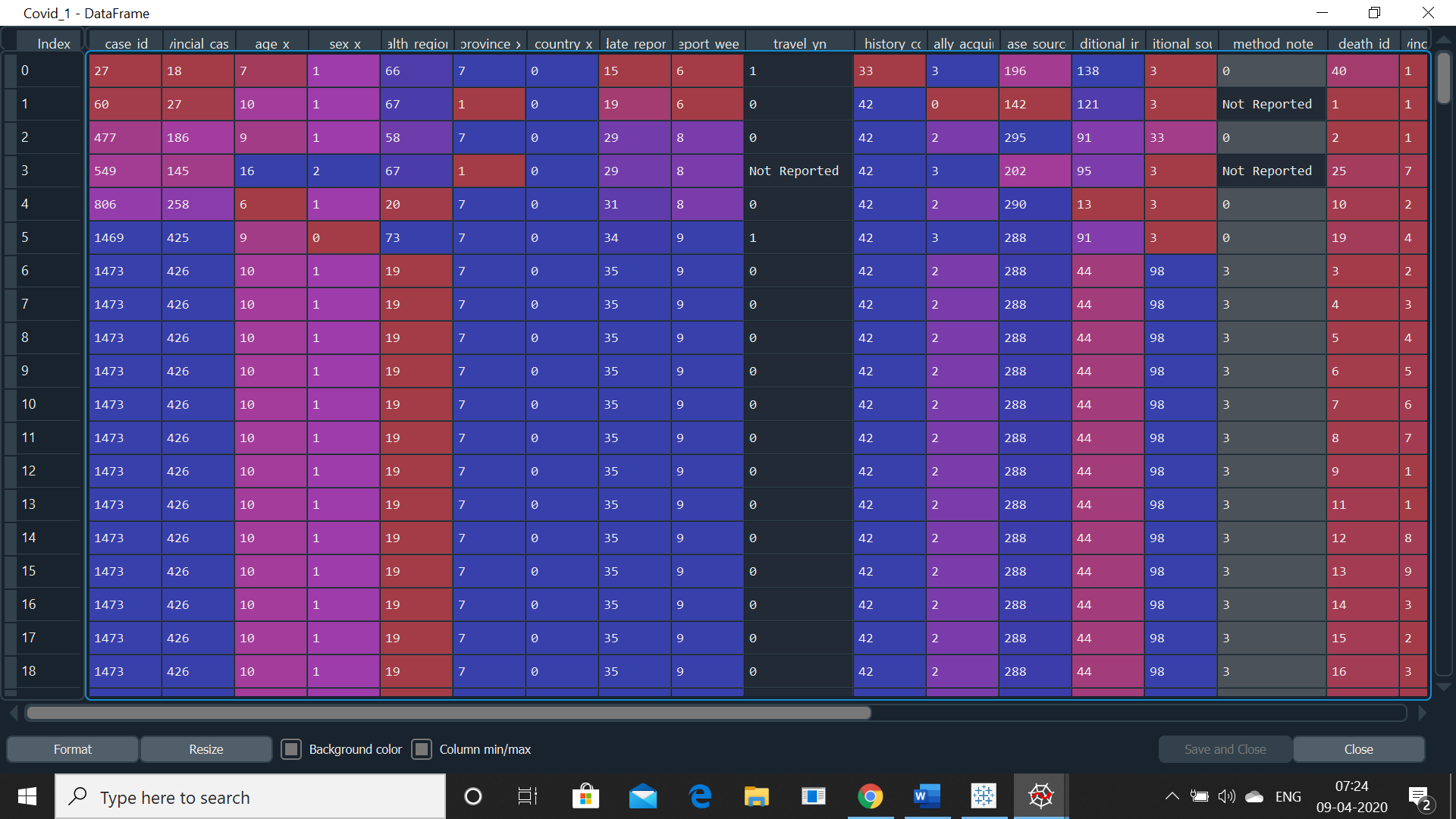


We remove the below columns as they were not relative to our findings:

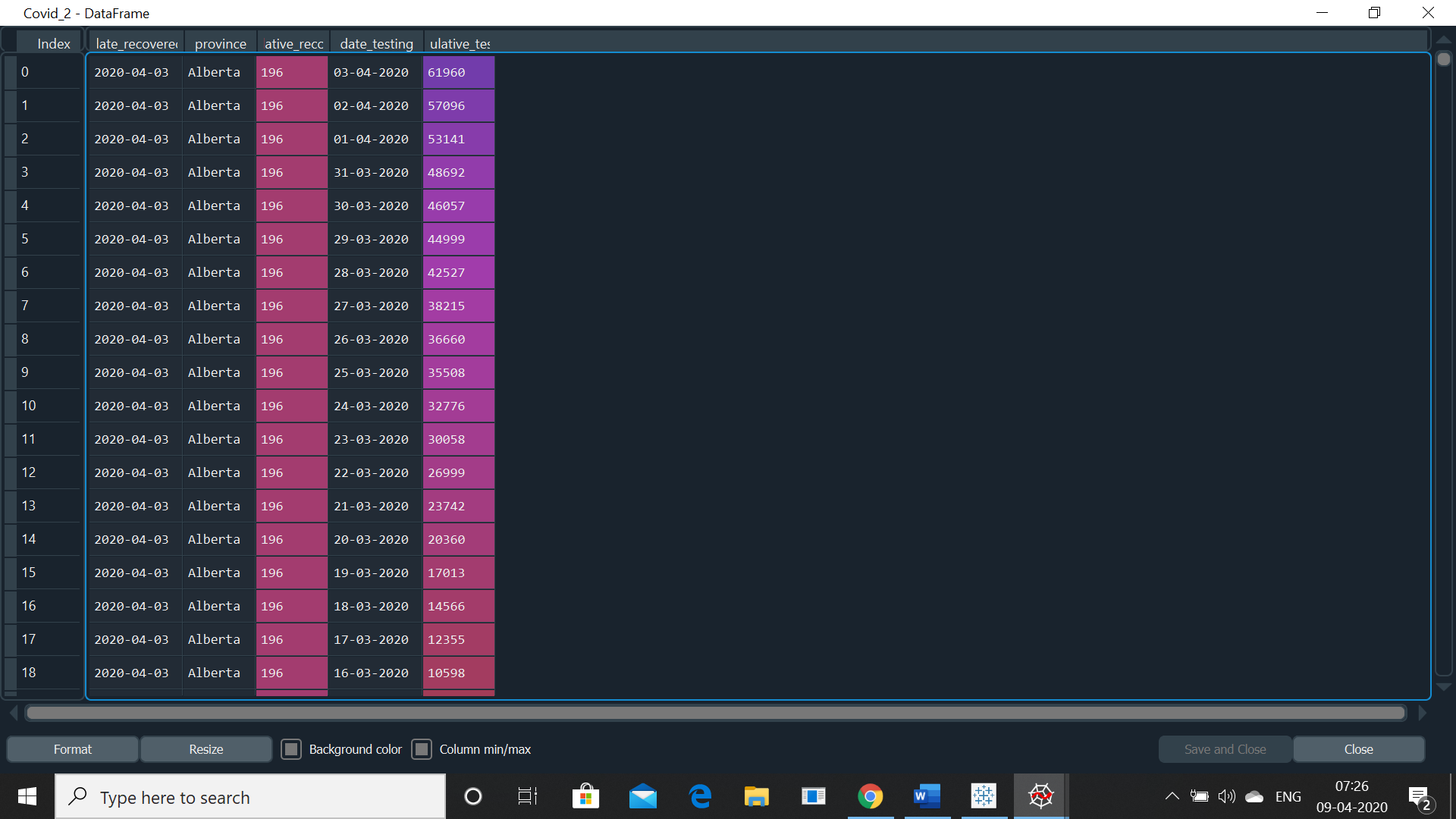
* additional\_info
* additional\_source
* province\_source
* source

**then we merge the data into 2 Dataframes:**

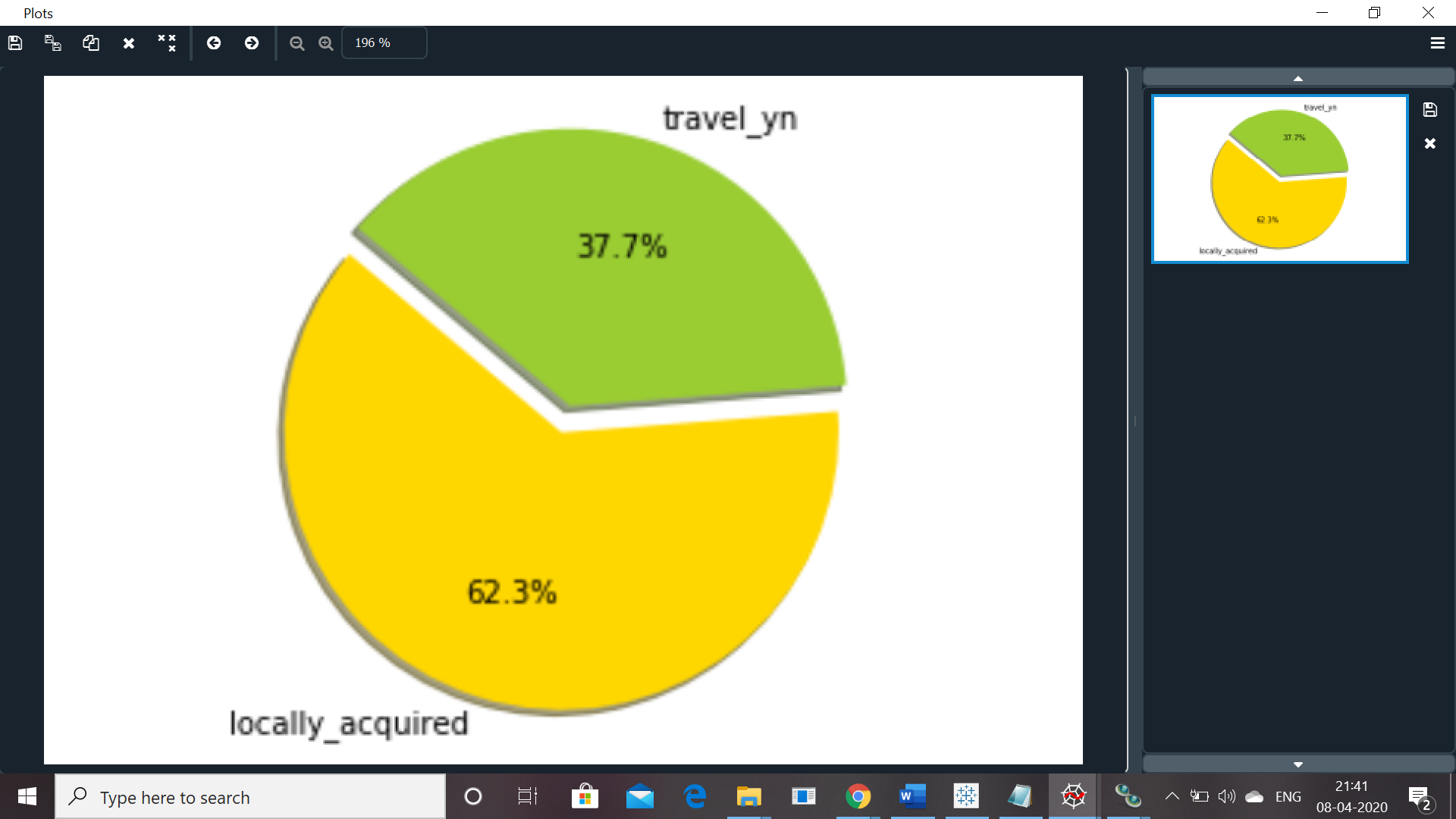
**Covid\_1: merger of cases and mortality**



**Covid\_2: merger of testing and recovered**



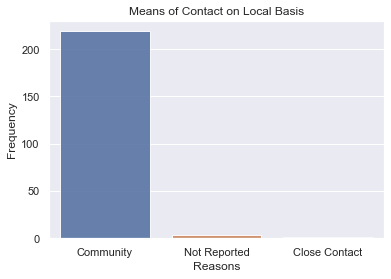
We plot a pie-chart to see how patients are infected



**Conclusion: Out of total number of patients 37.7% acquired the virus while travelling abroad, but 62.3% patients acquired it locally. i.e. directly or indirectly coming in contact with the patients who already had the virus.**

After, the above conclusion it is neccasary for us to find out how the majority of people came in contact with patients.

Thus, we plot the graph leading to the means of contact.

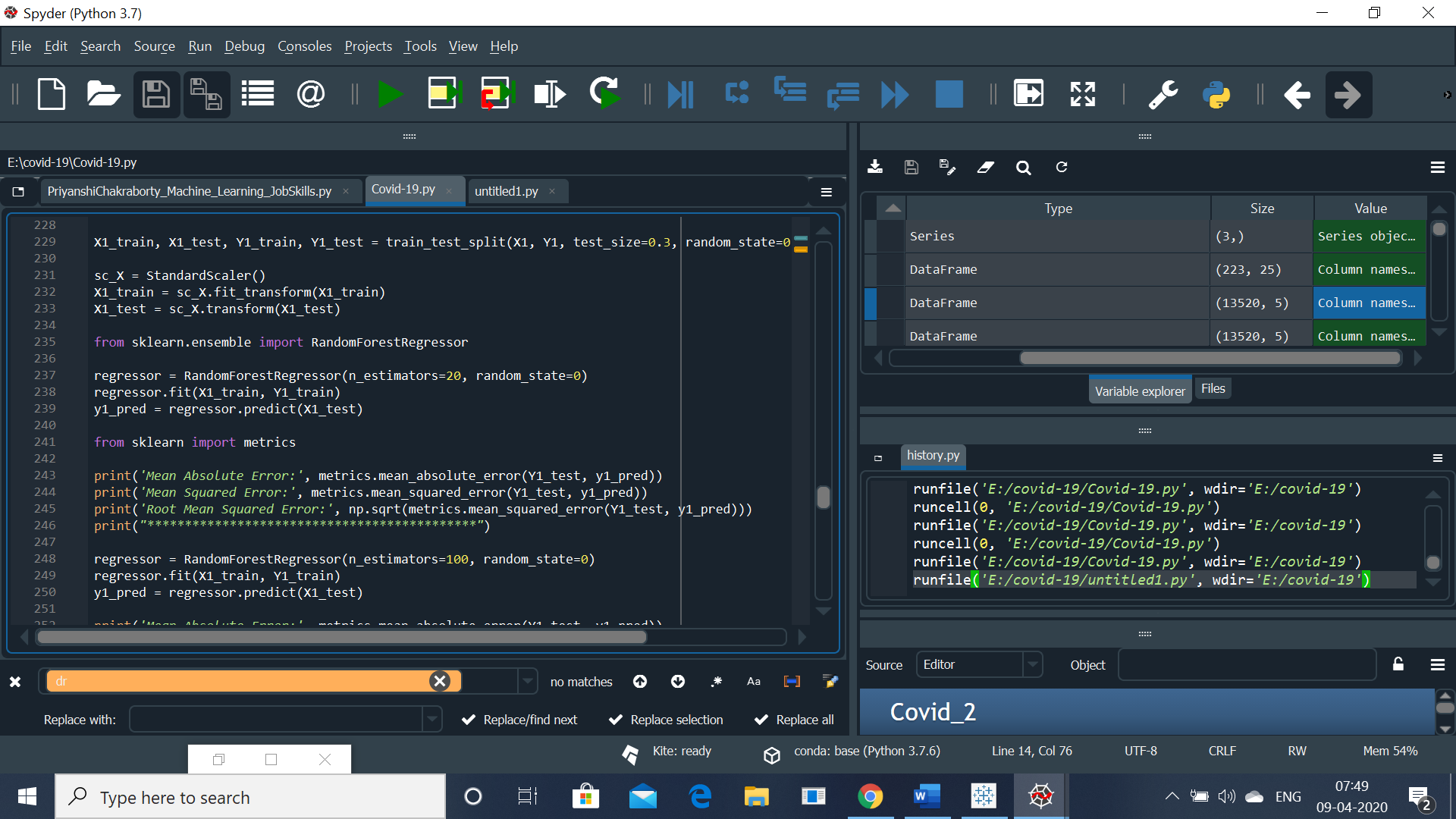


**Conclusion: After studying the above graph, we can come to the conclusion that the majority of the affected patients were actually not in close contact. Thus, we can assume that virus can be spread without any direct or physical contact between 2 people making it more deadly and easy to spread.**

**Machine Algorithm: Random Forest**

Random forest, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. The fundamental concept behind random forest is **“A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.”**

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



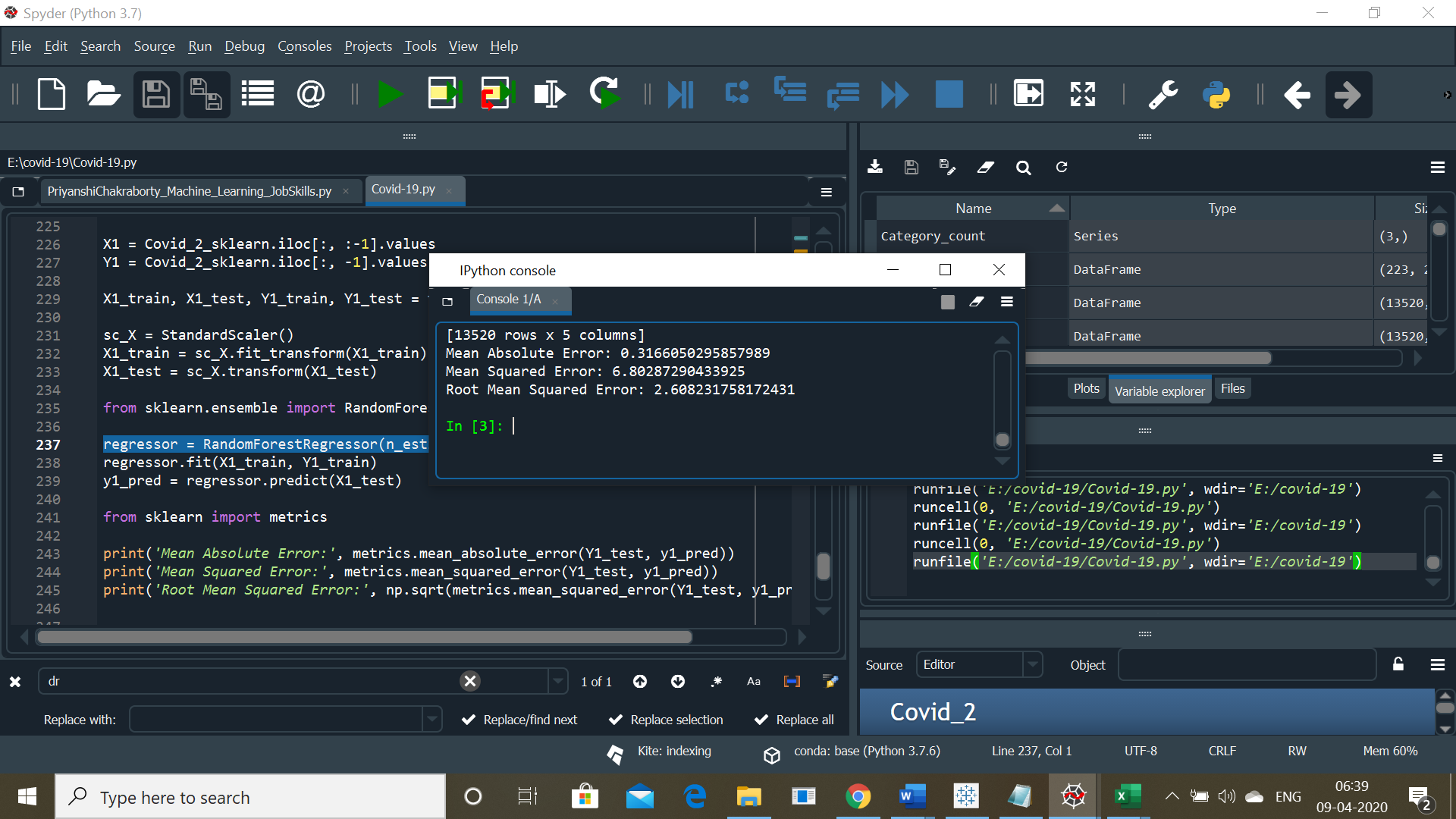
We applied Random Forest algorithm to our data Covid\_2. Assuming, the number of decision trees is equal to n\_estimators. Starting with 20 we increase the value of n\_estimators to 2000 and calculated the Mean Absolute Error, Mean Squared Error and Root Mean Squared Error (RMS) for each value of n\_estimator.

**regressor = RandomForestRegressor(n\_estimators=20, random\_state=0)**

**Mean Absolute Error: 0.3166050295857989**

**Mean Squared Error: 6.80287290433925**

**Root Mean Squared Error: 2.608231758172431**

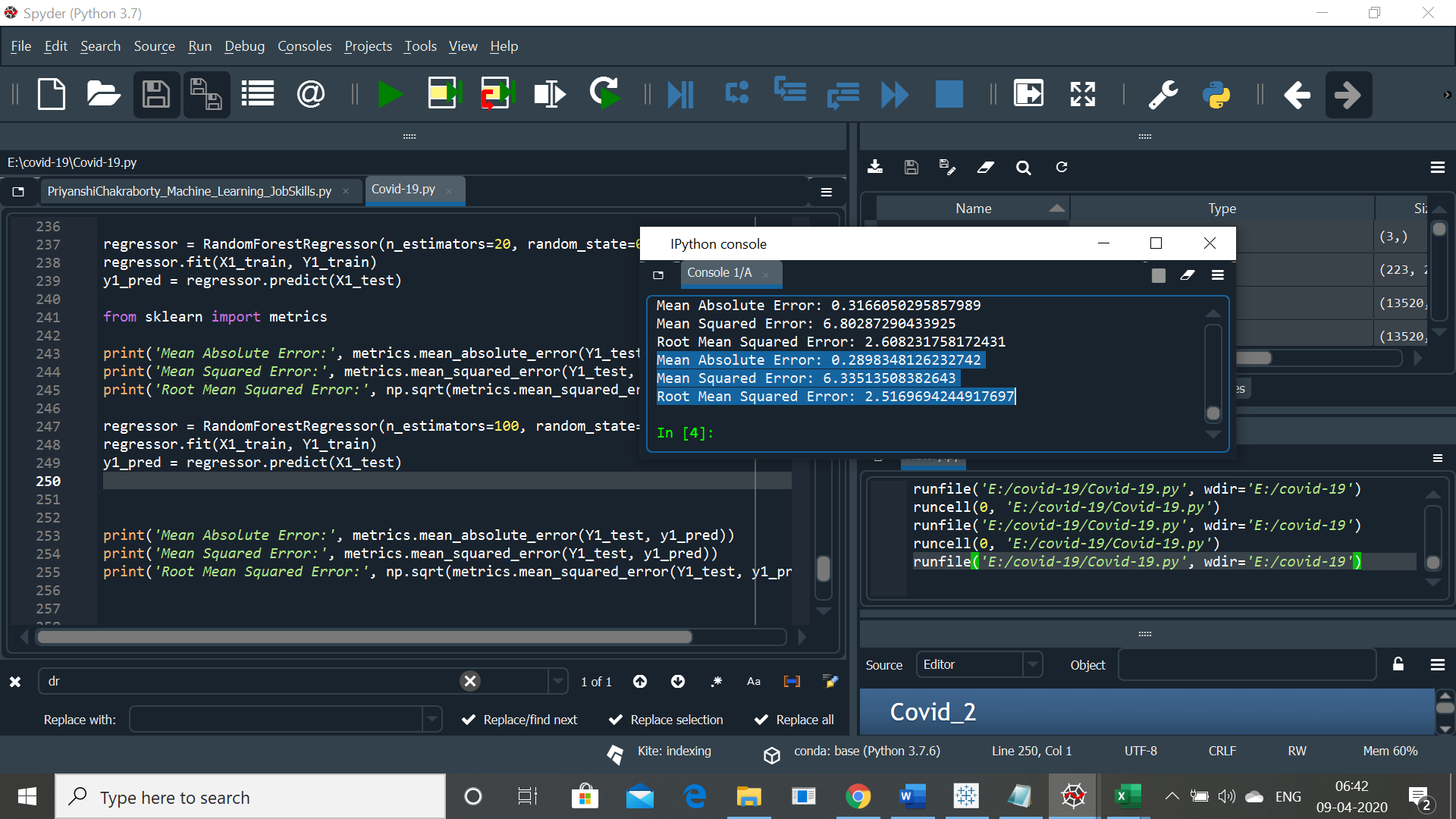


**regressor = RandomForestRegressor(n\_estimators=100, random\_state=0)**

**Mean Absolute Error: 0.2898348126232742**

**Mean Squared Error: 6.33513508382643**

**Root Mean Squared Error: 2.5169694244917697**

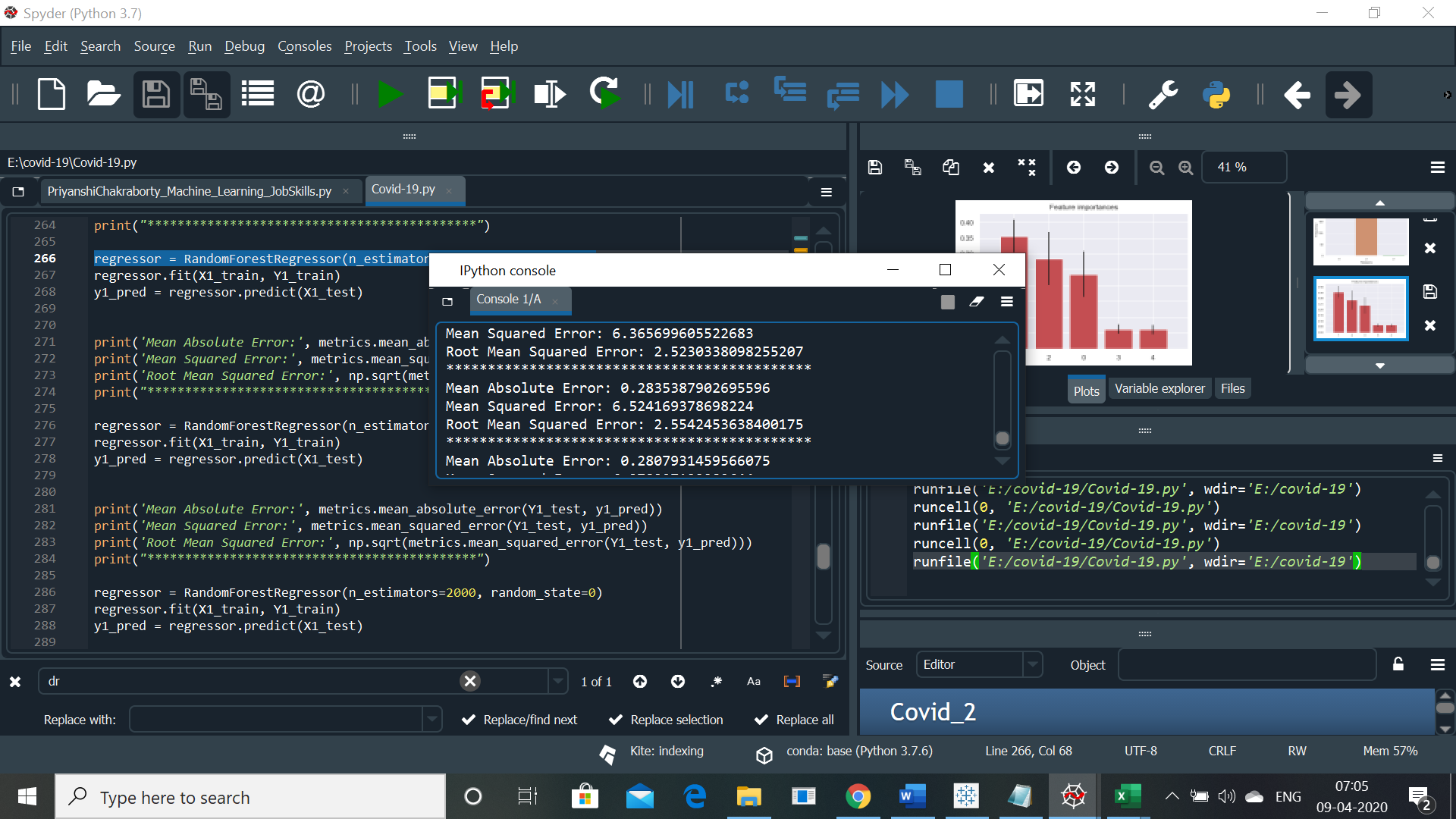


**regressor = RandomForestRegressor(n\_estimators=300, random\_state=0)**

**Mean Absolute Error: 0.2835387902695596**

**Mean Squared Error: 6.524169378698224**

**Root Mean Squared Error: 2.5542453638400175**

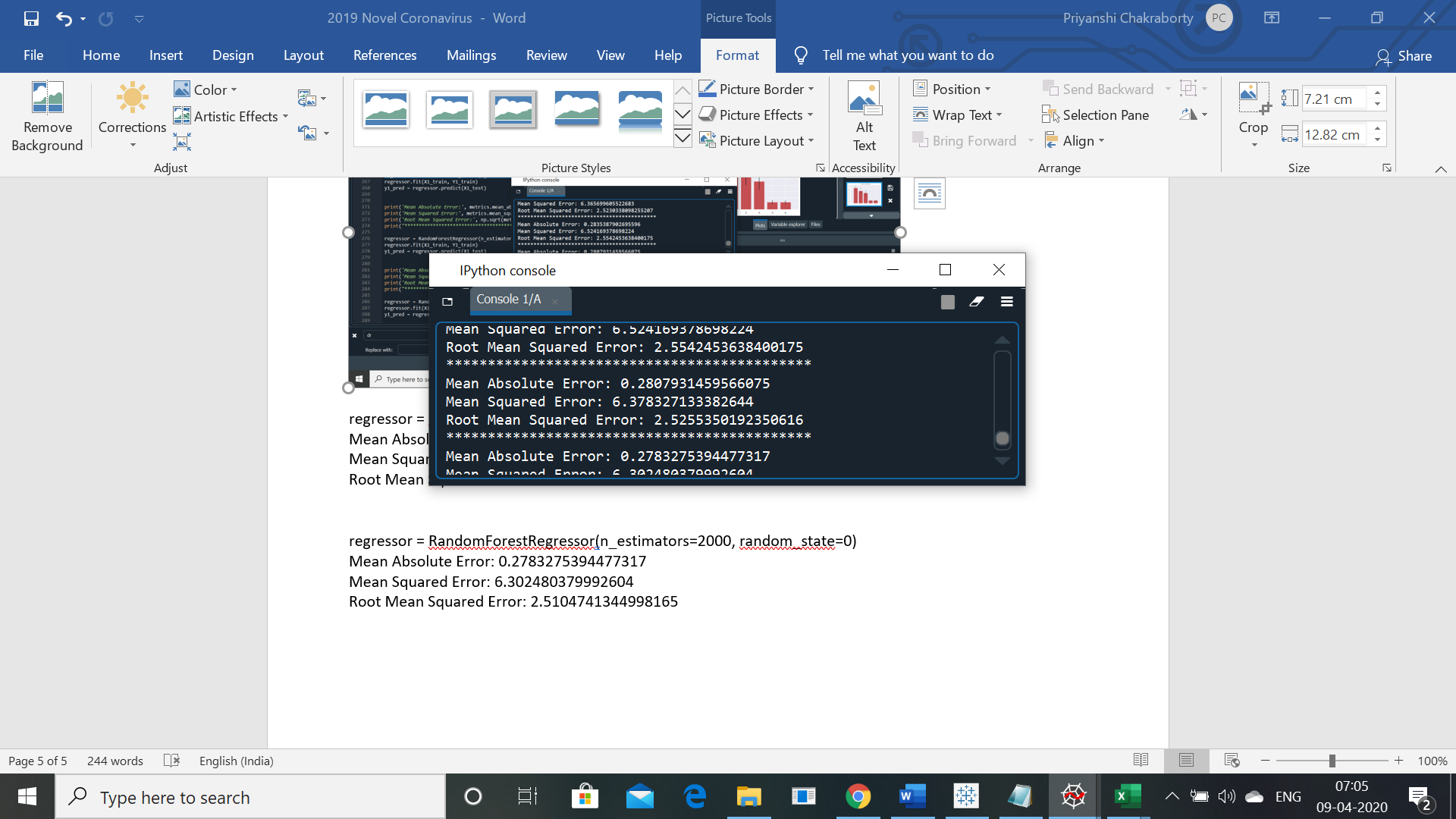


**regressor = RandomForestRegressor(n\_estimators=1000, random\_state=0)**

**Mean Absolute Error: 0.2807931459566075**

**Mean Squared Error: 6.378327133382644**

**Root Mean Squared Error: 2.5255350192350616**

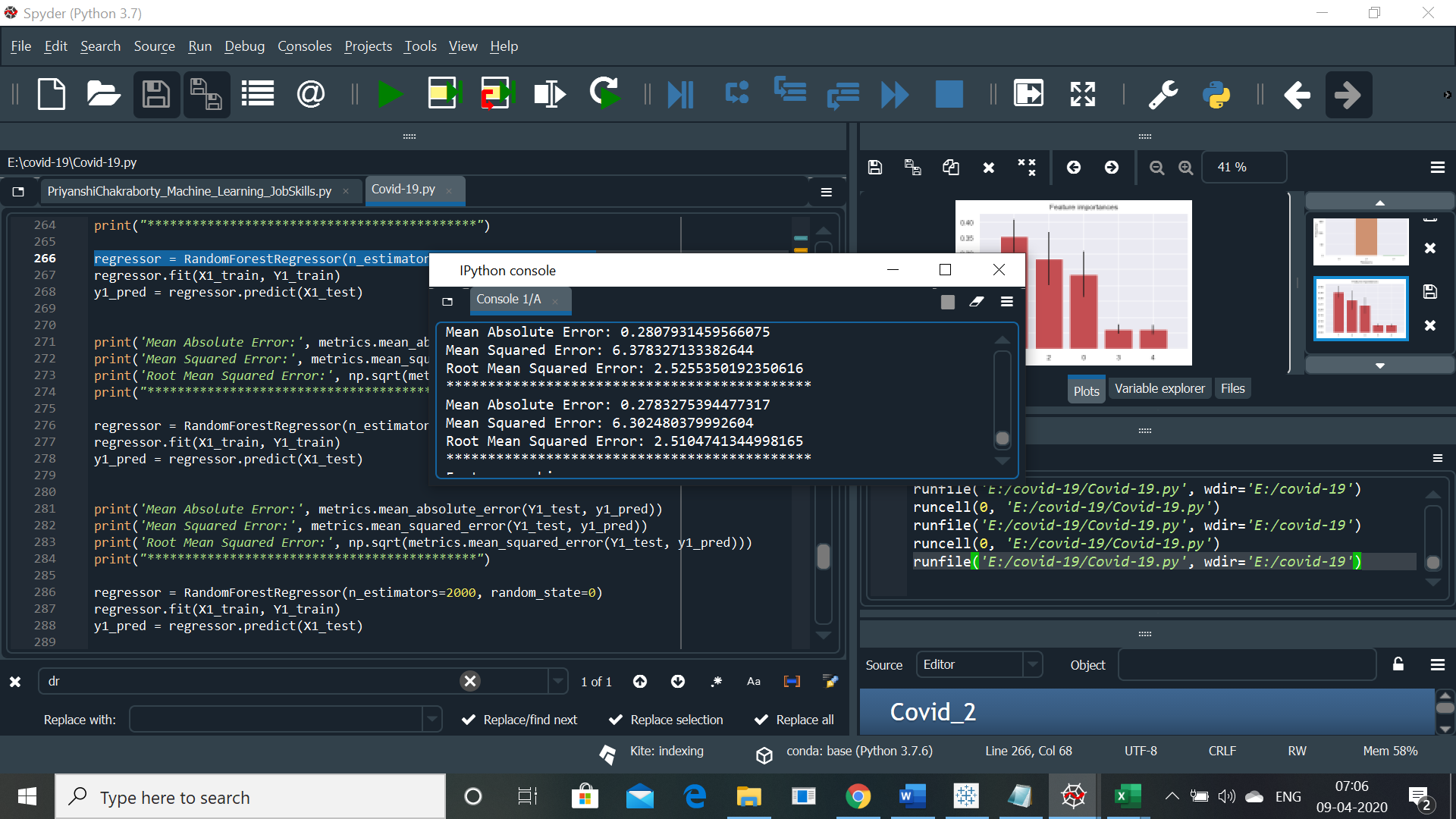


**regressor = RandomForestRegressor(n\_estimators=2000, random\_state=0)**

**Mean Absolute Error: 0.2783275394477317**

**Mean Squared Error: 6.302480379992604**

**Root Mean Squared Error: 2.5104741344998165**



**Conclusion: After going through the results , it becomes very clear that first when value of n\_estimator rises from 20 ,initially the value of RMS also increases, but after a point when value of n\_estimator reaches 1000 the value of RMS becomes almost equal to the value of RMS when n\_estimator was equal to 20.Futher when we increase the n\_estimator the value pf RMS keeps decreasing. This shows that we are reaching the number of decision trees which will give the most precise output.**

Now, let’s see which feature has major contribution in the output. We know that :

**Feature 0: date\_recovered**

**Feature 1: province**

**Feature 2: cumulative\_recovered**

**Feature 3: date\_testing**

**Feature 4: cumulative\_testing**

**Feature ranking:**

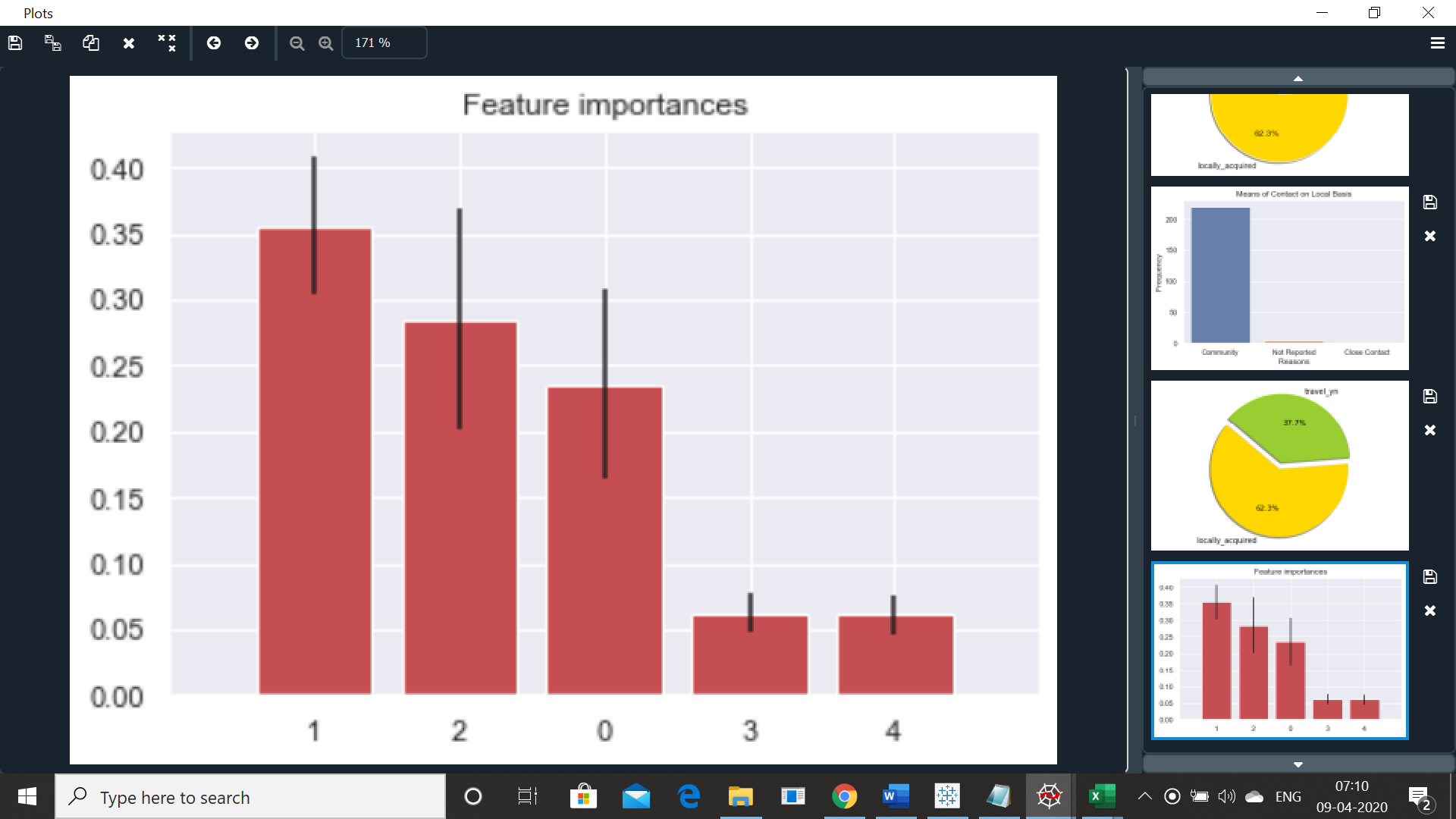
1. feature 1 (0.355632)

2. feature 2 (0.285299)

3. feature 0 (0.235688)

4. feature 3 (0.062131)

5. feature 4 (0.061250)



**Conclusion: We see the major contribution is from Feature 1: Province thus, we can conclude that the probability of getting infected is more in the states where there are already large number of people are infected.**

**Second major contribution, is of Cumulative\_recovered thus, we can safely assume that the number of Patients recovering is high thus, the treatment protocol is going in right direction.**