[Email address]

**ETL Project  
   
impact of covid-19 on Canadian employments**

**By Stephen and Latyr**

Table of Contents

[1 System Architecture 2](#_Toc80084115)

[1.1 Web 2](#_Toc80084116)

[1.2 Csv + Zip (inputs) 2](#_Toc80084117)

[1.3 Pandas 2](#_Toc80084118)

[1.4 Csv outputs 2](#_Toc80084119)

[1.5 PgAdmin 2](#_Toc80084120)

[2 Extract 3](#_Toc80084121)

[2.1 Kaggle’s COVID-19 dataset 3](#_Toc80084122)

[2.2 Statistics Canada 4](#_Toc80084123)

[3 Transform 6](#_Toc80084124)

[3.1 Kaggle’s COVID-19 dataset 6](#_Toc80084125)

[3.2 Statistics Canada 6](#_Toc80084126)

[4 Load 7](#_Toc80084127)

[4.1 Kaggle’s COVID-19 dataset 7](#_Toc80084128)

[4.2 Statistics Canada 7](#_Toc80084129)

As the current coronavirus which is also known at COVID-19 continuously spread across the world, we conducted this ETL project to allow people to assess its impact on labour market in Canada. To create this dataset, we utilized two sources including Kaggle’s COVID-19 dataset and employment data from StatsCan. These are rough overview and links of data involved:

To achieve our aim of creating such dataset, we have created programming codes designed to fulfill three steps of the ETL process with automation.

# System Architecture

The file below shows the different systems are interacting.

Diagram

Description automatically generated

## Web

The web was our main source of information.

## Csv + Zip (inputs)

Most files were available as csv files. In order to submit, we deleted big files that are not allowed by github.

## Pandas

Pandas was an familiar way of dealing with unstructured data.

## Csv outputs

These files will then be imported into our SQL database.

## PgAdmin

Our choice to use PgAdmin is based n the fact that the data must have a structure and a specific format. There SQL was the best option. Queries can be done using Date columns for inner joins.

# Extract

Extract data from designated source of this project. Two sources involved in the projects including Kaggle’s COVID-19 dataset and employment data from StatsCan. These are rough overview and links of data involved:

## Kaggle’s COVID-19 dataset

Link: <https://www.kaggle.com/ryanxjhan/coronaviruscovid19-canada#cases.csv>

A screenshot of a social media post

Description automatically generated

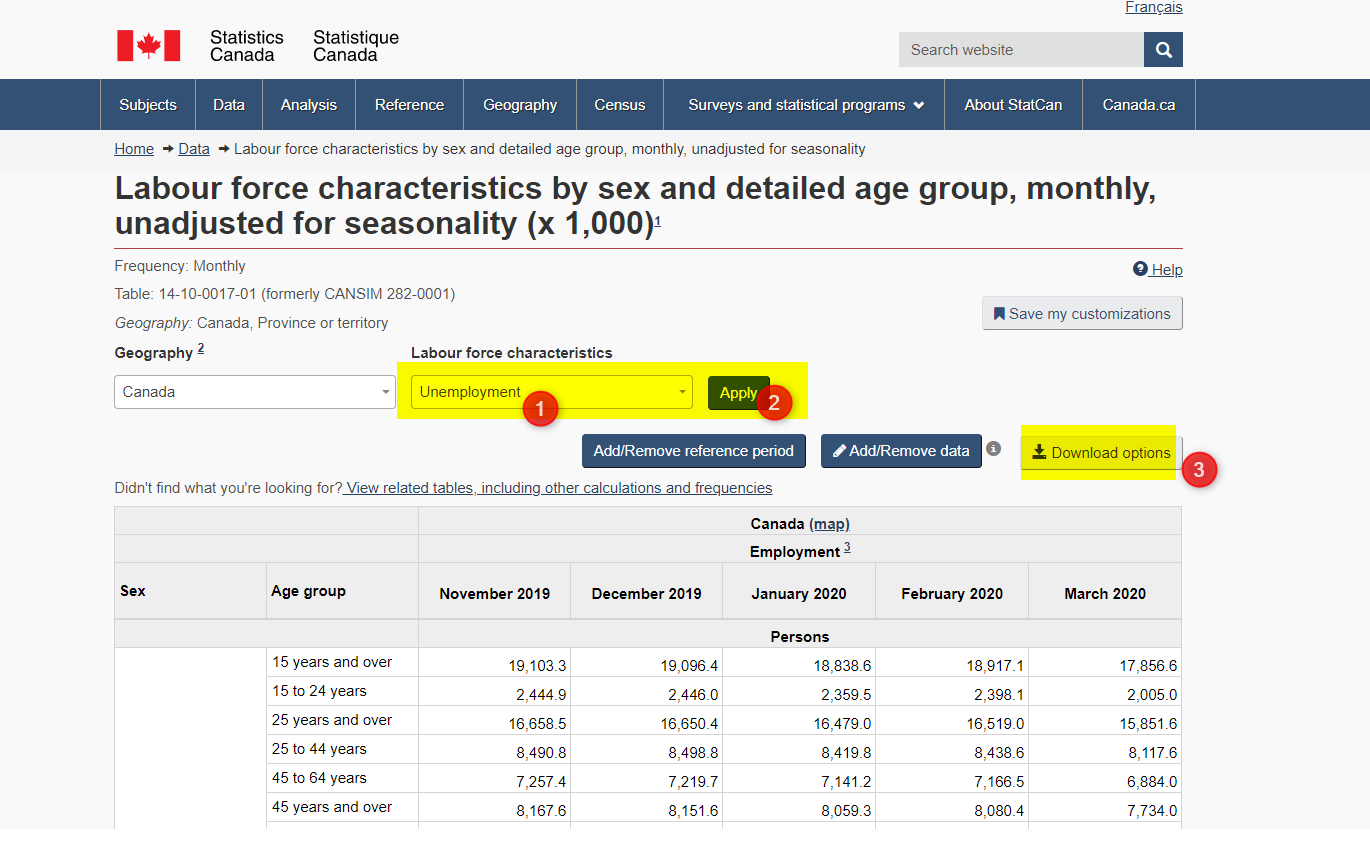
-This dataset is in the form of csv. After download and using python code to unzip the zip file, there will be 4 csv dataset concerning confirmed cases, mortality, recovered cases and testing.

## Statistics Canada

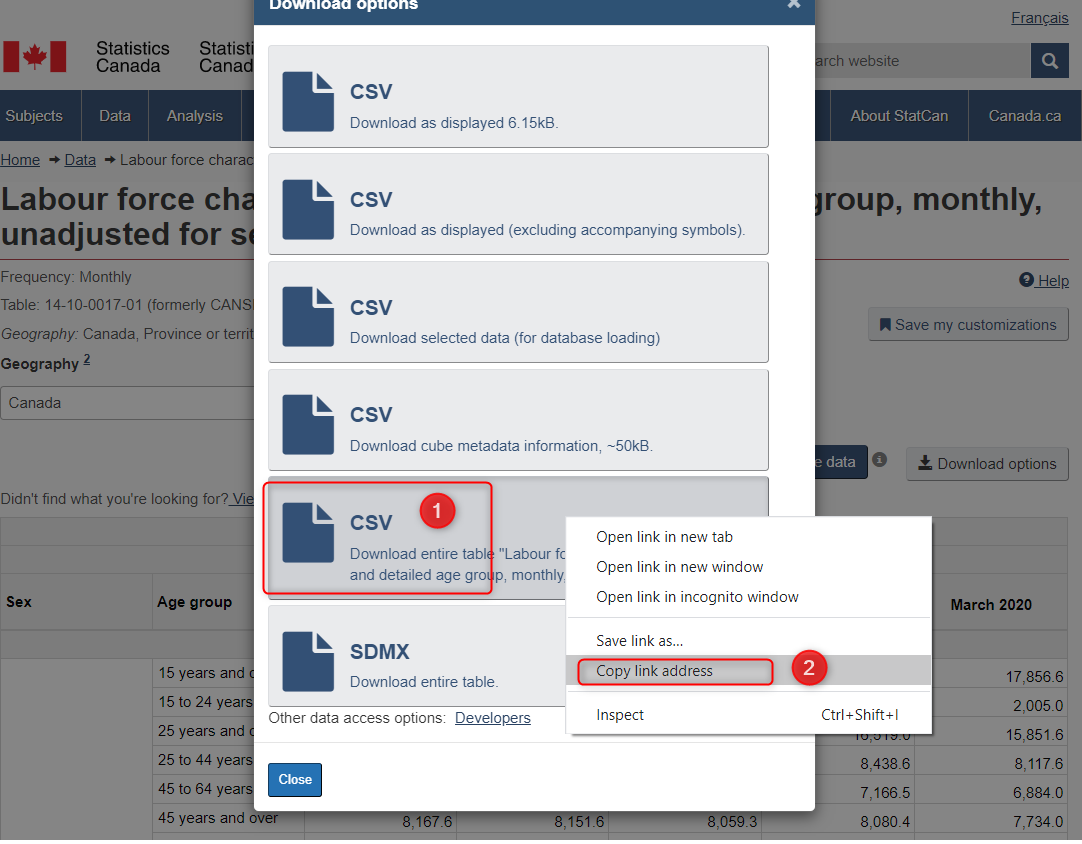
Visit the url below:

<https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410001701&pickMembers%5B0%5D=1.1&pickMembers%5B1%5D=2.3>

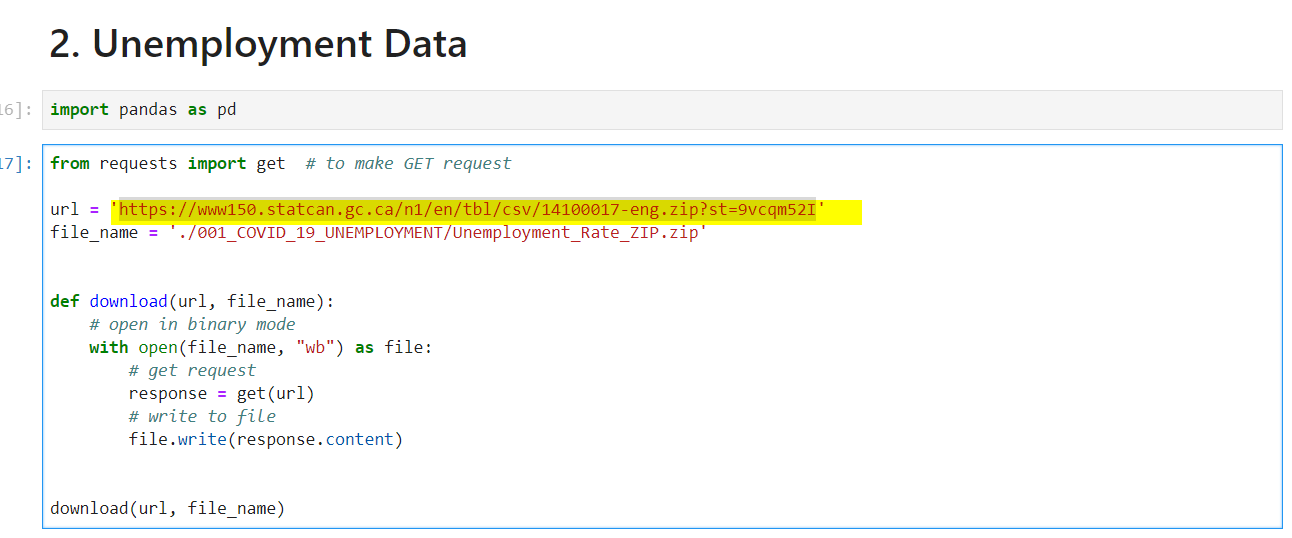
Then filter on the “unemployment” and “apply”, then click “download”

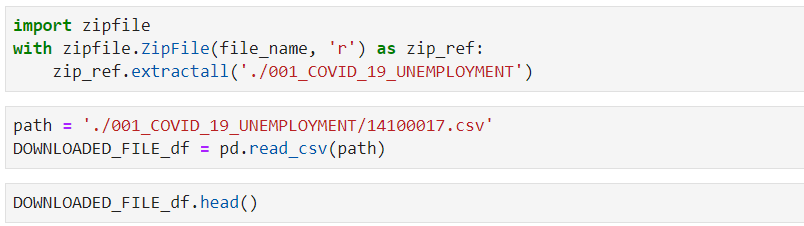


Select the second last and copy the link



Paste the link to your Jupyter notebook cell





This will download the CSV file and allow users to read the csv file by using Python.

# Transform

## Kaggle’s COVID-19 dataset

Due to many reasons including potentially incomplete report by patients and nature of survey, extensive data cleaning and transformation is required to handle missing data and combine similar responses. Following approach was conducted to transform data into more usable forms:

-Utilize fillna function in pandas to add appropriate value to cells with missing data through consideration of columns involved

-Many responses to question regarding if patients were locally infected and how they were infected have similar meanings, therefore, df.replace function is deployed to unify them to same values

-In order ensure existence of common column between employment dataset and COVID-19 dataset, values in common columns including sex and date are syncretized

-Groupby is used to generate table presenting counts of confirmed cases, mortality, recovered cases and testing. Appropriate groupby variables including date, sex and province are used to reflect important information and create common columns with the employment datasets.

-Any data concerning counts are set to integer.

## Statistics Canada

The csv file was imported into a dataframe and Python was used to accomplish the following steps:

Data types analysis

Narrowed down the columns to be used

Updated the Date format from ‘YYYY-MM’ to ‘YYYY-MM-DD’

Then filtered out all irrelevant dates to focus on the time frame of the pandemic (from November 2019 to March 2020).

Dropped all NaN values

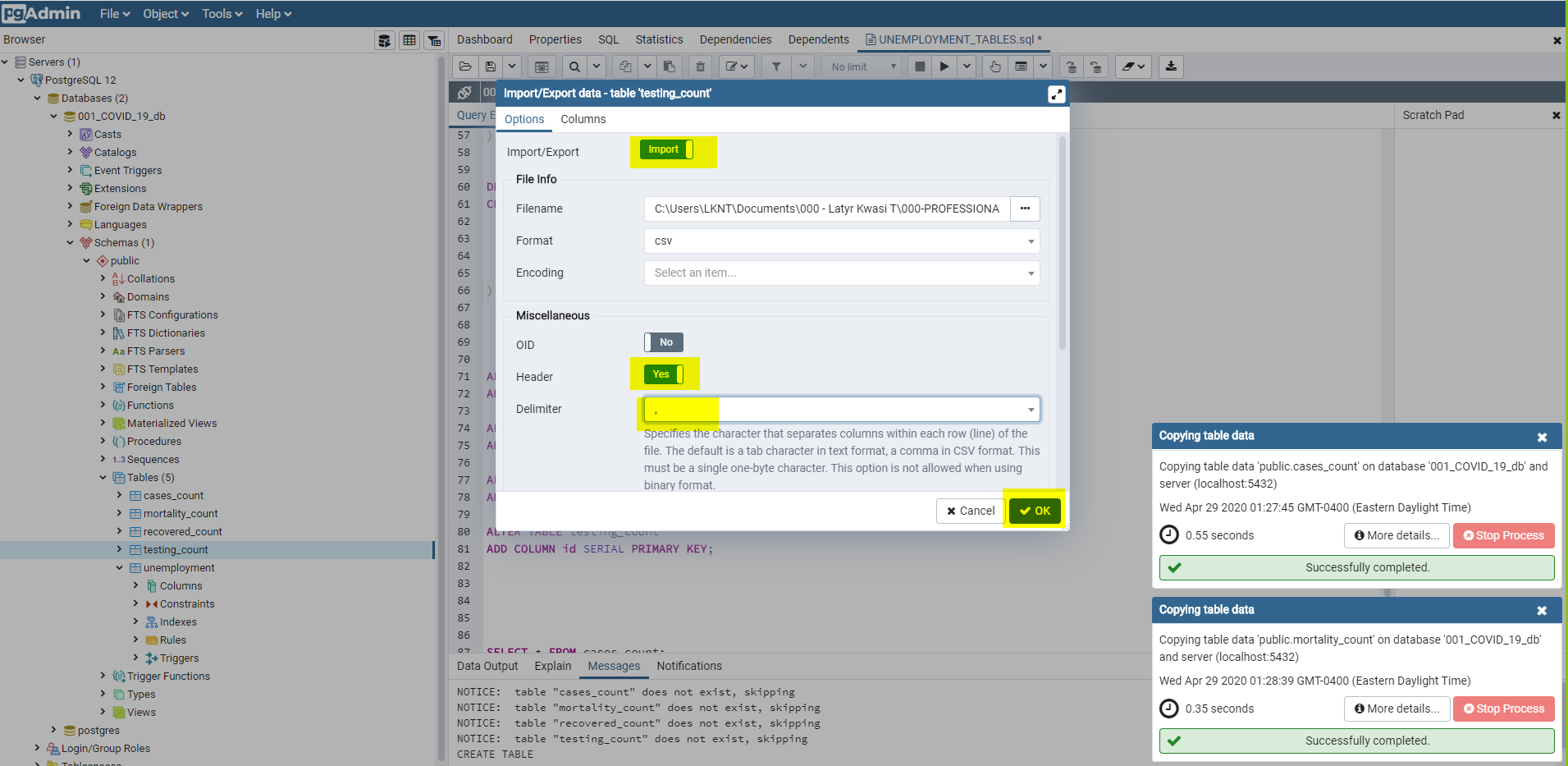
Exported the dataframe as a CSV file to be used in PgAdmin

# Load

## Kaggle’s COVID-19 dataset

After comprehensive cleaning and transforming of the COVID-19 data, summarized tables generated are uploaded to established SQL data base.

Pgadmin4 is chosen due to its ability to host permanent dataset. Additionally, with SQL server, users will be able to conduct meaningful queries and further data transformation through tools functions including JOIN, WHERE and others.



**Please use UTF8 (encoding), otherwise an error message will show.**

## Statistics Canada

* Created a database
* Created the table with related fields and types.
* Added primary keys
* Data import