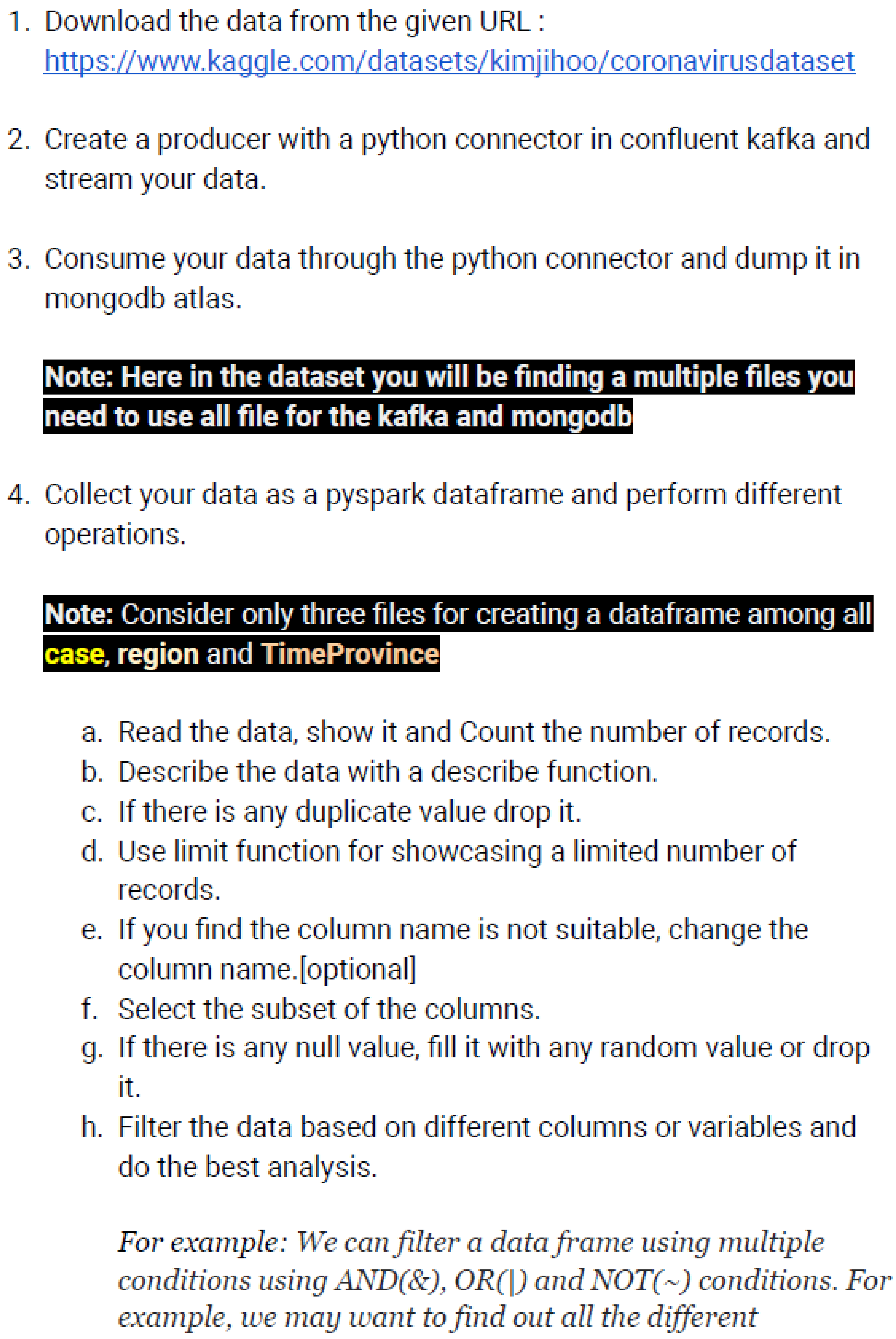
I have completed the assignment as per below:



1. Download the data from the given URL : https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset

2. Create a producer with a python connector in confluent kafka and stream your data.

3. Consume your data through the python connector and dump it in mongodb atlas.

   Note: Here in the dataset you will be finding a multiple files you need to use all file for the kafka and mongodb

4. Collect your data as a pyspark dataframe and perform different operations.

   Note: Consider only three files for creating a dataframe among all **case**, **region** and **TimeProvince**

   a. Read the data, show it and Count the number of records.
   b. Describe the data with a describe function.
   c. If there is any duplicate value drop it.
   d. Use limit function for showcasing a limited number of records.
   e. If you find the column name is not suitable, change the column name.[optional]
   f. Select the subset of the columns.
   g. If there is any null value, fill it with any random value or drop it.
   h. Filter the data based on different columns or variables and do the best analysis.

      For example: We can filter a data frame using multiple conditions using AND(&), OR(|) and NOT(~) conditions. For example, we may want to find out all the different