Instruction for the Project Final Submission:

# Dataset:

Refer to the following resources:

1. [Novel Corona Virus 2019](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset) for dataset description

2. [Johns Hopkins Coronavirus Resource Center](https://coronavirus.jhu.edu/map.html) for graphical view of the data

Please note that you’ll be using time series data to make predictions, please use data only until 27th May 2020 (inclusive). The above-mentioned dataset also contains dedicated files for USA, they will not be required for this project and can be ignored.

# Language and Tools:

Python programming language should be used for this project. Allowed machine learning libraries include:

* Pytorch
* Scikit-Learn
* Pandas
* Numpy

You are allowed to use any library for visualization of data, if necessary. Some good candidates are matplotlib, seaborn and ggplot.

Note: Please include any more information you may need

# Project Task:

You will only be forecasting for Pakistan only in this project.

1. Load and process the concerned data, you will use **time\_series\_covid\_19\_confirmed**
2. Identify 5 most related countries that have a similar trend of infection as Pakistan. While comparing trends please keep in mind the 27th May restriction on data.
3. Use the data of these countries to predict the spread of Coronavirus in Pakistan. Your prediction must start from 10th May and until 27th June so you will also be predicting trend for one month into the future (31 days into future).

*Hint: While comparing trends and predicting, use number of days versus amount of cases as a measure of spread, instead of date against amount of cases*

1. Your expected output will be a list of numbers where each next individual element refers to the total **new** cases reported on that day.
2. Use the prediction from 10th May until 27th May and report how well the model performed.
3. Now use Pakistan time series and predict future cases until 27th June (one month into the future)
4. Plot the original time series and predicted time series and compare their difference/similarities explain any peculiarities if found. Also explain briefly how different countries predict different trends for Pakistan and what do you think is the best one.

# Important Guidelines:

1. There are a plenty of forecasting models you can use:
   1. Autoregression
   2. Moving Average
   3. ARIMA
   4. RNN/LSTM

You should try as many as you like and go with what works best for you. You can also introduce more complex ideas into your models. Sometimes implementing ideas based on intuition goes a long way.

You can also use other models beyond the ones listed above, however you must implement all models yourself. Plagiarism will not be tolerated.

1. You are restricted to Pytorch and Scikit-Learn for implementation of your models.
2. You are restricted to the dataset mentioned above, you can utilize all the files in it however you please, using any other data source other than specified is not allowed.
3. You cannot use any third-party libraries and tools other than the one mentioned. You can however use Rapid Miner for your own personal experimentation.

# Deliverables:

1. Python Jupyter file containing all written code
2. A project report in PDF format explaining your approach, results and findings.