**Q1**

**The approach of the problem:**

* Population of India
* Urban population (covid hotspots are for most cities)
* % Of urban poor (cannot afford the medical facilities and hence need help)
* % Lower and middle middle class (recently lost jobs and would be using money to prioritise the basis need hence cannot afford to buy medical supplies
* Cost of each kit and items in each kit

**Data:**

* Pop = 1.4 billion
* % Urban population = 33%
* % Of urban poor = 25%
* % Of middle class = 28% (we will take this as 20 % as the other 8 % would be in upper middle class)

**Calculation:**

The population that needs to be delivered with the kit = 1.4 billion x 33/100 x (25+20)/100 = 207900000

That is 20.8 lakh kits need to be distributed

**Cost of one kit:**

* Kit items: 3 sanitizers, 15(N95) facemasks, medication amounting to Rs 50
* Therefore, the cost will be 3 x 30 + 15\* 20rs + 50 = Rs 440 (approx. cost of the kit meant to last for 1 month)
* Total cost = 20.8lakh x 230 = 9150 crore rupees

**Time required:**

* Places where these kits are manufactured: Hyderabad, Pune (pharmaceutical companies) and the distance from these places to all the metro cities can be reached within 54hrs by a freight train.
* Time required for manufacturing of each kit can be calculated from the time of mask production (slowest among all the other contents) = 3lac masks per day, we need 15\*20.8 masks in total therefore the time required for mask production = 104 days from start to finish
* Total time required to finish the task from day of approval will be close to 120days (manufacturing+ transportation +distribution)

1. SOLUTION (approx.):
2. 20.8 lakh kits
3. 9150 crore rupees
4. 120 days

**Q3**

**Project name: Diamond price prediction**

**Link:** **https://www.kaggle.com/code/gravindernathyadav/diamonds**

The aim of this project is to create a machine learning model which has a high accuracy in predicting the prices. This is a common project done by data scientists whilst learning and hence I had taken up the project. The data was readily available on Kaggle. The data contains various parameters that may or may not affect the price of the diamond these were namely carat, cut, colour, clarity, depth, table, x, y, z and at last price.

The roadmap was as follows:

1. Clean the given data and search for missing values

There were values of x, y, z which were equal to 0 and since these occurred very less, hence dropped

1. Plot graphs to understand the relationship between parameters and the price

The price highly depends upon carat, x, y, z but not so much on table and dept). The colour, clarity and the cut too play a role in price(minor)

1. Removing outliers using inter quartile range
2. Feature selection

After hot encoding of the data in colour, cut, clarity the features considered are namely: colour, cut, clarity, carat, x, y, z

1. Model train, prediction, and scoring

Linear Regression model was chosen for its simplicity. The data was split into train and test data and the features were scaled using the StandardScaler package in sklearn. The model was fit onto training data and the prediction was calculated. R2 score was calculated = 0.913, that is 91.3% accuracy.