The Competition “Analysis: COVID-19 patient recovery data”:

This challenge is based on a dataset of 1085 patients who were tested positive for Coronavirus disease (COVID-19) across different countries. The data also has details on patients who were able to recover from the disease along with patients who were not able to survive the outbreak. The challenge is to understand the data at hand, deep dive into the information that is provided and try to uncover patterns. As part of the competition, we have raised some questions which you have to answer. In addition to that, you are urged to answer some questions of your own and share those with us. In the end, prepare a presentation (not more than 10 slides) where the synthesis of the entire analysis is presented. Do remember to share along your jupyter notebooks in which you will execute your entire analysis.

Important Note: This is a dataset sourced from the Open access epidemiological data from the COVID-19 outbreak. The COVID-19 outbreak is a global pandemic and a serious situation and any insight that you may get from this competition may help address some questions about the COVID-19, but please do remember that those insights and facts may only be applicable to the sample at hand.

Problem Description:

On 31 December 2019, WHO was informed of cases of pneumonia of unknown cause in Wuhan City, China. A novel coronavirus was identified as the cause by Chinese authorities on 7 January 2020 and was temporarily named “2019-nCoV”. On 11th March 2020, the WHO Director-General declared the novel coronavirus (2019-nCoV) outbreak as a global pandemic. The number of confirmed cases worldwide has exceeded 2,00,000. It took over three months to reach the first 1,00,00 confirmed cases, and only 12 days to reach the next 1,00,000.

This competition is about exploring the sample data on COVID-19. The **COVID-19** dataset has 1085 rows and 14 columns. This is a sample of 1085 patients who have been tested Positive for Coronavirus disease (COVID-19). You have to give insights out of the given information.

Data dictionary:

1. **case_in_country:** Country-wise occurrence of the case. Like-7 means 7th patient in a particular country.
2. **country:** Name of the patient's country.
3. **location:** Location where the patient is admitted in the hospital.
4. **gender:** Gender of the patient(male/female).
5. **age:** Age of the patient in years.
6. **visiting Wuhan:** Has the patient visited Wuhan in recent times ("0": no, "1": yes).
7. **from Wuhan:** Is the patient a resident of Wuhan ("0": no, "1": yes).
8. **death:** Is the patient dead or alive? ("0": alive, "1": dead).
9. **recovered:** Has the patient recovered from the disease? ("0”: not, "1": yes).
10. **symptom_onset:** Date on which symptoms started to appear.
11. **hosp_visit_date:** Date on which the patient visited the hospital for the first time.
12. **exposure_end:** Date on which exposure to an infected person/infected place ended.
13. **symptom:** Symptoms like cough, fever, etc.
14. **summary:** Summary of the patient from which all this information is extracted.

Analysis to be done: -

- Carry out descriptive statistics and share your findings.
- Check the data for missing values and outliers. Comment.
- Do univariate, bivariate and multivariate analysis. Share your inferences.
- In the exploration of the data, try to create several questions and hypothesis of your own. Answer those questions and test your hypothesis based on the available data. Some examples for your perusal:

  - You would have read so much about coronavirus being deadly for older people. Can you validate this hypothesis from the data you have available with you?
  - Are females getting more infected then males?
  - What is the proportion of cases across countries?
  - Which symptoms are most prominent in the patients?
  - Is there any pattern between early hospital visits, symptom_onset, and recovery rate? (Do not limit yourself to just the pointers above, try to explore data extensively to **discover patterns** and **insights**).
- Can you try building a classification model to predict whether the patient will recover or not? (Target variable – recovered) If you go ahead with the model-building exercise, validate your results with different model performance measures.

RULES OF THE COMPETITION:

1. Upload a Presentation which showcase your insights from the analysis. Elaborate it with graphs and plots. The maximum limit for slides is 10. Also, upload the R/Python code notebook where you have done the analysis.
2. Top performers will be those who answered not only the above questions *but also executed* ***out of the box*** *to present the analysis. (Use all the knowledge, commands and techniques learned so far).*
3. For assessing top performers, more weightage will be given to those who go the extra mile and make unique contributions to their analysis.

Best of Luck and Happy Learning!