PROJECT DESCRIPTION

COVID-19 pandemic has created a havoc for over a year and a half now. The government and WHO has jointly decided to perform some analysis over the situation to plan over for the further steps for prevention. The government has hired some analysts and agents to create a scheme for the insurance and even understand the strategy that can be implemented over various states or across the nation unanimously.

AIM/Problem Statement: The approach is to **Analyse – Predict – Conclude**

- To Analyse the COVID-19 Globally and Nationally based on the datasets provided to you.
- To Predict the insurance amount based on the inputs provided by the data.
- Perform Statistical analysis over all the datasets, with valid reasoning.
- Provide conclusive steps to prevent COVID-19 spread, as the fear of the third wave emerges with an analytical approach.

Approach: Three datasets "Covid_19_clean_complete", "Statewise Testing Details", "Insurance" is provided.

- **For the Statewise Testing Details**: Perform statistical data analysis, show the correlation between the columns with respect to cases. In case if any columns are categorical, use as. factor(), or any other methods to convert the data into numbers.

    1. Provide correlation tests on various columns with respect to positive tested patients.
    2. Perform a sum of cases with respect to states
    3. Provide the average daily positive test using formula: Sum of positive tests/Number of positive tests observations
    4. Provided the average daily negative test using formula: Sum of negative tests/Number of negative tests observations.
- **For Covid 19 Clean Complete/Covid 19**: Perform regional analysis over the active cases based on WHO region.
    1. As the data is quite inconsistent as we observe that the country has reported multiple instances of cases over a single day, perform a data cleaning by taking the sum of cases with respect to dates.
    2. Post Data Cleaning perform seasonality analysis (line chart) to understand whether there was any seasonality observed with the rise and fall of the cases. Perform this with respect to WHO Regions.
    3. With respect to the WHO region:
        I. Perform correlation test:
            - Average Deaths to Average Recovery
            - Average Active to Average Deaths
        II. Take any WHO region and compute the hypothesis that the mean recovery rate for any specific date is similar to the mean recovery rate of any previous period.
        III. Test the hypothesis that the Mean deaths in the month of March 2020 is greater than the Mean deaths in the month of September 2020, take any WHO region.
        IV. Perform a linear model (lm) to predict (pick the data from Jan 2020 – Sept 2020) the number of deaths for the month of October 2020.

V. Show using a pie chart with respect to WHO regions:
   - Deaths
   - Recovery
   - Active
   - Confirmed

VI. Analyse the variance for the number of cases in the month of May with respect to the number of cases in the month of October.

VII. How well are Confirmed and Active cases correlated?

- **For Insurance**: Various columns in the insurance file are mentioned. The government ran a pilot program to sell the insurance based on various factors.
  1. Based on regions which region would be best suitable to target and sell the insurance.
  2. Should this insurance be charged on a similar basis to that of a smoker to non-smoker? If yes, please justify, if no please justify.
  3. Perform a linear model with charges as a response variable and age, bmi, smoker as predictor variables. This is a pricing model which the analyst must propose to the company and so to the government.
  4. Perform correlation test on
     - Charges ~ Age
     - Charges ~ BMI
     - Charges ~ children
     - Charges ~ Smoker
     - Charges ~ Region
  5. Create a pricing model from the above linear model which would provide a price/charges which an individual must pay based on their characteristics like age, bmi and smoker.

- **For Non-Life Insurance:** What are additional or continuous efforts that the Company should keep on tracking to ensure they are constantly updated and re-aligned to new reality. *This may also include your suggestions on further analysis a Company should carry out to be ahead of competitions.*
  E.g: Continue monitoring of actual vs. expected number of COVID cases and re-align the model

  Also, you are required to share the final recommendations and action plan you propose for the Company. This should include actionable in terms of Products, Risk, Underwriting, Reinsurance and continuous analysis performance.

  Eg: Additional risk factors the Company should consider in post COVID scenario.
  E.g: Additional Reinsurance program Company should opt for to tackle Pandemic.
  E.g. Specific measures for states with very high positivity rate

CONCLUDE:

- What necessary steps should specific countries take to avoid further spread of the COVID

- Which regions (WHO Regions) should be more careful

- Propose your pricing model, how justified is it to use only these 3 parameters

- Is it good to pay charges for the insurance?

Note:

1. Use graphs wherever necessary.
2. Perform data cleaning wherever necessary.
3. The 3 csv files are totally independent.
4. Make use of appropriate statistical methods wherever necessary.