Final Analytics' Steps

1. Log on to NYU dumbo and connect to Hive with the following command.
   a) Beeline
   b) !connect jdbc:hive2://babar.es.its.nyu.edu:10000/
   c) USE NetID;
   d) SHOW TABLES;
2. Now create two tables from the cleaned data, one for covid infection rate and one for happiness score.
   a) create external table covid (country string, confirmed int, death int, recovered int, active int, death_rate double, recovered_rate double, one_week_increase int, one_week_increase_rate double)
   row format delimited fields terminated by ','
   location '/user/NetID/part-r-00000';
   b) create external table happiness (country string, happiness_score double, gdp_per_capita double, social_support double, life_expectancy double, freedom double, generosity double, government_trust double)
   row format delimited fields terminated by ','
   location '/user/NetID/pbdaa_project/Input1/';
3. Close the current PUTTY terminal and open a Windows CMD, type in the command "ssh -L 4483:babar.es.its.nyu.edu:10000 NetID@dumbo.hpc.nyu.edu"
4. Open Tableau and make sure the required ClouderaHive ODBC driver is installed.
5. Connect to Cloudera Hadoop with the server as "localhost", port as "4483", type as "HiveServer2", Authentication way as "Username and Password", Transport as "SASL", and your NetID and corresponding password.
6. Type in your NetID in "Schema" to load your tables.
7. Type in "covid" and "happiness" to load the two tables.
8. Click "New Custom SQL".
9. Type in the following SQL query
   a) Select c.country, c.confirmed, c.recovered_rate, h.life_expectancy, h.freedom, h.generosity, h.government_trust

      From NetID.covid c
      Inner join NetID.happiness h on c.country = h.country

10. Now we are just plotting the with different combinations of factors. Open a worksheet and add longitude and latitude to column and row respectively. Drag the column "country" from the resulting table given the SQL query to the Detail under "Marks". Now a world map with all countries having available data should be appeared.
    a) To visualize the correlation between confirmed cases and life expectancy of each country, use "sum(life_expectancy)" to determine the color and "sum(confirmed)" to determine the size of the circle. Here I use 8 levels of Orange-Blue Diverging and

reverse the palette so that higher the magnitude of life expectancy turns to be orange. Also adjust the size of all circles proportionally by clicking the size button so that it provides a clear view over the map.

b) To visualize the correlation between recovered rate and life expectancy of each country, use “sum(life_expectancy)” to determine the color and “sum(recovered_rate)” to determine the size of the circle. Here I use 8 levels of Orange-Blue Diverging and reverse the palette so that higher the magnitude of life expectancy turns to be orange. Also adjust the size of all circles proportionally by clicking the size button so that it provides a clear view over the map.

c) Now for the remaining three factors: freedom, generosity, and government_trust, just repeat steps a and b for all three factors.

d) So we get 8 plots, showing the correlations (or not) between confirmed cases and freedom, generosity, life_expectancy, government_trust and the correlation between recovered rate and freedom, generosity, life_expectancy, and government_trust respectively around the world. From the plots we may inspect and analyze if certain happiness factors has any correlation with the COVID-19 confirmed cases and recovered rate nationally.