Test

We are pleased to invite you to the interview process for our Data Science Team! This is a practical exercise that will test your programming and analytical skills, please **include your codes as a PDF** in the submission. The programming language that is acceptable is python or R.

Instructions: Please read carefully

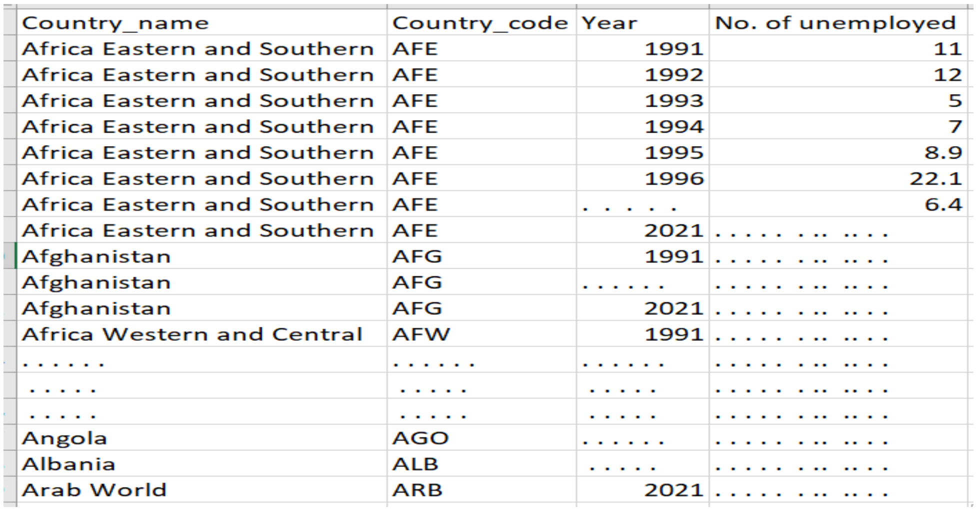
- ❖ **Submit 1 pdf file with all the answers. The submitted pdf file name should be in '<your_full_name>_<date>.pdf' format.**
- ❖ **Your code, comments & output should be present in the pdf. Please make sure that all the output code and text are organized and readable in the submitted PDF.**
- ❖ You may not consult with any other person regarding the test.
- ❖ You may use internet searches, books, or notes you have on hand.
- ❖ The test has 7 parts, **all of which are mandatory**. Failing to complete any one part would result in the rejection of the submission.
- ❖ In case of doubts please make thoughtful assumptions.

Part 0: Reading the data

- Please find the data (unemployment_analysis.csv) and take it as the input (as data frame).
- For all columns (having years), convert the strings into floats. (for e.g., 6.8 lacs □ 6.8)
- Display the column names along with their datatypes.
- Generate the unique country names along with their corresponding country code.
- Find total unemployed people for all the years for each country.

Part 1: Data cleaning

- Write a function called data_cleaning() which, when called, would perform the following activity:
  1. Find for any missing value in all the columns, display them.
     If any missing value exists, then replace them with the average of the corresponding country. Then, again, check for null values.
  2. For the countries 'Benin', 'Bahrain', find if any outliers exist. If yes, replace them with mean/median/mode.
  3. Create a new column, named, "Year", which would have all the years as per each country & beside that column, add a new one named, "No. of unemployed", which would have the corresponding total values.
  4. Change the column name "Country Name" to "Country_name" & "Country Code" to "Country_code".
     Finally, this is how our data, after the changes, should look like: (this is only for reference)
     [Let's call it df_pivot]



Country_name	Country_code	Year	No. of unemployed
Africa Eastern and Southern	AFE	1991	11
Africa Eastern and Southern	AFE	1992	12
Africa Eastern and Southern	AFE	1993	5
Africa Eastern and Southern	AFE	1994	7
Africa Eastern and Southern	AFE	1995	8.9
Africa Eastern and Southern	AFE	1996	22.1
Africa Eastern and Southern	AFE	. . . . .	6.4
Africa Eastern and Southern	AFE	2021	. . . .
Afghanistan	AFG	1991	. . . .
Afghanistan	AFG	. . . . . .	. . . .
Afghanistan	AFG	2021	. . . .
Africa Western and Central	AFW	1991	. . . .
. . . . . .	. . . . . .	. . . . . .	. . . .
. . . . .	. . . . .	. . . . .	. . . .
. . . . .	. . . . .	. . . . .	. . . .
Angola	AGO	. . . . . .	. . . .
Albania	ALB	. . . . .	. . . .
Arab World	ARB	2021	. . . .

- Write a function called descriptive_stats(country_code) which would –
    1. Give the mean, median, mode & standard deviation for the parametrized country over all the years. [e.g., descriptive_stats('WSM') would give the descriptive statistics of this country.]
    2. Give the year during which the country's (of the passed country_code) unemployment was minimum & maximum.
    3. Find the top 5 countries which had maximum unemployment in 2021.
    4. Find the top 3 countries that had unemployment greater than 5 lacs in the year 2021.
    5. Calculate the change (in percentage) the countries (from ques. 3) saw in unemployment starting from 1991 to 2021 (year on year). [for e.g., change_1 = (1991_value – 1992_value.)/1991_value, change_2 = (1992_val – 1993_val)/1992_val & so on.]

Part 3: Prescriptive statistics

- From the questions in descriptive_statistics – part, answer the following:
    1. After getting statistics for a country over all years, can you tell in which year the country having country_code 'BGR' had the minimum unemployment issue?
    2. Create a new dataframe which would give us the country names along with their corresponding country codes. We would also have a new column along each country which would give us the increase or decrease percentage in unemployment that each country saw from 1991 to 2021.
    3. Can you compare between the minimum unemployment seen over all years for the country 'Japan', with the value seen in the previous & next 1 year of that year? Write a brief note on the same.
    4. Among 'MDA', 'NAC', 'PAN', 'PAK', 'UGA' which countries saw a huge jump in the unemployment numbers from 2019 to 2021, in both upward & downward direction separately?
    5. What would you say about the change in percentage seen from 1991 to 2021 for the country 'LSO'? Was it a predictable upward/downward movement? [To see that, you may have to plot the numbers for all the years for 'LSO'.]

Part 4: Simple Machine learning questions

- Write a function called predict_future('how_many_consecutive_future_year_values_you_need') - that would –
    1. Predict the next 2 years' values for all the countries. [i.e., for 2022 & 2023]
       [Use 2 models – Moving Average (take 3 months – for e.g., for 2005 = avg of (2002, 2003, 2004)) & ARIMA – for forecasting. Also, compare the MAPE of the 2 models.]
    2. Plot a line graph that would have the actual values we have in the dataset along with the forecasted values.
    3. For each year out of the 2 years (2022, 2023), find the maximum & minimum number of unemployed for each country.
    4. Find which forecasted year has the maximum overall unemployment.
    5. Find which countries' unemployment in 2022 [which you'll be forecasting] increased in comparison to 2021.

Part 5: Visualization

- Write a code to display the following graphs: (make sure the graphs are labelled properly.)
    1. A bar graph plotting the top 10 countries' in unemployment in the year 2021.
    2. A line graph showing the percentage change in unemployment from 2000 to 2019 (year on year) for the top 5 countries that had maximum unemployment in 2021.

3. A bar graph that would plot the top 3 countries in unemployment in 2021 along with their forecast in 2022.
4. For the country code ‘WSM’, plot a line graph that would show us how the unemployment changed every 5 years. (like in 1991, 1996, ….)
5. Please add any insights you could derive from all the graphs above.

Part 6: About the Previous projects

- Please describe any interesting project you did in the Data Science domain in more than 300 words. Attach Github links if possible.

Part 7: Time management

- Can you please share your thoughts, in less than 120 words, on “If you get selected, how will you manage your time for this full-time internship opportunity”

Best of luck!