

Assignment 2 - Pandas

Name :NAZRIN , Surname: THANIKATTIL RAFEEQUE, Student ID : 21031314

The Story

How many people are unemployed now? How much is the labour force either employed or unemployed and actively look for the jobs? Does gender is a factor that decides the employability of any individual? Above all, how these constrains are interconnected to the health of an economy? All these conceptions are statistically analysed and discussed.

This Analysis construe that Netherlands seems to have the highest labour force of 80.31% with India and Sri Lanka being the least number of 57.93 and 59.86% of labour force with the total population. Netherlands and Canada have the highest female labour force of 75.75 %and 75.13 %with the last being Pakistan 24.24%. The Highest rate of Unemployment among Females and Females are at Spain and Italy (about 50% of total female population). and Lowest is at Pakistan

Unemployment ratio to the total labour force in countries are analysed to see that Spain and Italy are having the highest unemployment rate and Pakistan and Japan has the lowest. From this analysis it can be incurred that gender is another factor that determines employability in few countries like Sri Lanka ,India and Italy female unemployment rate is quiet high comparing with males.

While taking the economy of different countries USA and India have highest GDP value .It can be seen that GDP and unemployment are correlated by 11.11 % negatively to infer that unemployment effects the economy of every nation adversely evaluation of different employment sectors prove that most of the highly GDP rated countries give immense opportunities in Service sector jobs rather than agriculture and industries and it is increasing rapidly over last 4 years. whereas the south Asian countries mainly employed in agricultural and industries. However in recent years the service sector growing at a very faster rate in the developing countries and is contributing a major share in terms of output, income and employment. Even the productivity per worker is becoming higher in-service sector when compared to agriculture and industrial sectors.

On making a region wise comparison it can be concluded that North American Countries Have the least Unemployment and Labour force Rate, wherein they maintain a 78 % of their total employability in service Sector and provide employability irrespective of their Gender and maintain the highest GDP rate as well. South Asian countries occupy the second level with Lower unemployment rate and high GDP by focussing on all employment sectors almost equally , however they fail to attain The gender equality in employment .At the same time the European countries interestingly have a high value of unemployment,but maintain gender equality and higher value of GDP by enhancing their employability in Service sectors.

Data Preparation

```
In [ ]: import pandas as pd
import numpy as np
import wbgapi as wb
```

Countries

```
In [ ]: """Chosen a list of 12 countries """

country_codes = ['GBR', 'ITA', 'NLD', 'ESP', 'JPN', 'MNG', 'KOR', 'IND', 'LKA', 'PAK', 'USA', 'CAN']
```

```
In [ ]: """A dictionary is created with country code to are assigned with country names"""

country_names = ({'GBR': 'United Kingdom', 'ITA': 'Italy', 'NLD': 'Netherlands',
                  'JPN': 'Japan', 'MNG': 'Mongolia', 'KOR': 'Korea',
                  'IND': 'India', 'LKA': 'Srilanka', 'PAK': 'Pakisthan',
                  'USA': 'United States', 'CAN': 'Canada'})
```

```
In [ ]: """A dictionary of country groups created to represent the 4 regions with keys
being Country codes and values being country code"""

country_groups = ({'GBR': 'Europe', 'ITA': 'Europe', 'NLD': 'Europe',
                  'ESP': 'Europe', 'JPN': 'East Asia', 'MNG': 'East Asia', 'KOR': 'East Asia',
                  'IND': 'South Asia', 'LKA': 'South Asia', 'PAK': 'South Asia',
                  'USA': 'North America', 'CAN': 'North America'})
```

Indicators

```
In [ ]: '''Search for indicators in umemployment in world bank data'''
wb.search("unemployment")
```

```
In [ ]: '''Description regarding the chosen indicators is searched'''

wb.series.info("SL.UEM.1524.FE.NE.ZS")
```

```
'''Defining the Indicator list 1 with desired Wbgapi ID's'''

indicator_ids_1 = ['SL.UEM.1524.FE.NE.ZS', 'SL.UEM.1524.MA.NE.ZS', 'SL.UEM.1524.NE.ZS',
                  'SL.TLF.ACTI.ZS', 'SL.TLF.ACTI.FE.ZS', 'SL.TLF.ACTI.MA.ZS']
```

```
In [ ]: wb.series.info('SL.UEM.1524.MA.NE.ZS')
```

```
In [ ]: wb.series.info('SL.UEM.1524.NE.ZS')
```

```
In [ ]: '''Search indicators for Labor Force in World Bank Data'''
        wb.search("Force")
```

```
In [ ]: '''Gathered information regarding the selected indicators'''

        wb.series.info('SL.TLF.ACTI.ZS')
        wb.series.info('SL.TLF.ACTI.FE.ZS')
        wb.series.info('SL.TLF.ACTI.MA.ZS')
```

```
In [ ]: '''Defining the Indicator list 1 with desired Wbgapi ID's'''

indicator_ids_1 = ['SL.UEM.1524.FE.NE.ZS', 'SL.UEM.1524.MA.NE.ZS', 'SL.UEM.1524.NE.ZS',
                  'SL.TLF.ACTI.ZS', 'SL.TLF.ACTI.FE.ZS', 'SL.TLF.ACTI.MA.ZS']
```

```
In [ ]: '''Searching for desired indicators for second dataframe'''

        wb.search("employment")
```

```
In [ ]: wb.search("gdp")
```

'''SL.IND.EMPL.MA.ZS : Employment in industry, male (% of male employment) (modeled ILO estimate) SL.SRV.EMPL.ZS : Employment in services (% of total employment) (modeled ILO estimate) SL.AGR.EMPL.ZS : Employment in agriculture (% of total employment) (modeled ILO estimate)'''

SL.IND.EMPL.FE.ZS : Employment in industry, female (% of female employment) (modeled ILO estimate) SL.IND.EMPL.ZS : Employment in industry (% of total employment) (modeled ILO estimate) NY.GDP.MKTP.CD : GDP (current US\$) '''

```
In [ ]: '''Defining the indicator List 2'''

indicator_ids_2 = ['NY.GDP.PCAP.KD', 'NY.GDP.MKTP.PP.CD', 'SL.SRV.EMPL.ZS', 'SL.AGR.EMPL.ZS', 'SL.IND.EMPL.ZS']
```

Data Frames

```
In [ ]: '''Defining the pandas DataFrame by Fetching the values of indicator series from wbapi along with time
series for last 10 years'''
df = wb.data.DataFrame(indicator_ids_1, country_codes, time=range(2008, 2019, 2), columns='series')
df
```

```
In [ ]: '''Defining the index names for the dataframe'''

df.index.names = ["Country", "Year"]
```

```
In [ ]: '''Performing multindexing on Dataframe columns & splits labor Force
and unemployment into level 0 and male female catogorisation as level1 columns'''

col = pd.MultiIndex(levels=[['labor force(%)', 'unemployment(%)'], ['Male', 'Female', 'Total']], codes=[[0, 0, 0, 1, 1, 1], [1, 0, 2, 1, 0, 2]])
```

```
In [ ]: '''Defining the multiindexed dataframe df_1, describing Labor force and unemployment for last 10 years'''

df.columns = col
df_1 = round(pd.DataFrame(df, columns=df.columns), 2)
df_1
```

```
In [ ]: '''Defining the second DataFrame by Fetching the values of indicator series from wbapi along with time
series from last 10 years'''

df2 = wb.data.DataFrame(indicator_ids_2, country_codes, time=range(2008, 2019, 2), columns='series')
```

```
In [ ]: '''Defining the index names'''

df2.index.names = ["Country", "Year"]
```

```
In [ ]: '''Defining the Multiindex DataFrame using list of desired Tuples in the indicator list the define the DataFrame'''

col = pd.MultiIndex.from_tuples([('GDP', 'gross'), ('GDP', 'PPP'), ('Employment', 'Agriculture'), ('Employment', 'industries'),
                                ('Employment', 'Services')], names=["C0", "C1"])

df2.columns = col
df_2 = round(pd.DataFrame(df2, columns=df2.columns), 2)
df_2
```

Data Analysis

```
In [ ]: #creating pivot table with total labor force and finding the maximum labour force in male and female

df_pivot = df_1.pivot_table(['labor force(%)'], index='Year', columns="Country")
df_max = df_pivot.max()
display(df_pivot, df_max)
```

Netherlands seems to have the highest labour force of 80.31% when comparing with other countries, with India and Srilanka being the least number of 57.93 and 59.86% of labour force with the total population. In this Japan have the highest male force of JPN - 86.47%. Netherlands and Canada have the highest female labor force of 75.75 and 75.13 with the last being Pakistan 24.24%.

```
In [ ]: #Slicing the dataframe for Labour Force and Unemployment data
idx = pd.IndexSlice
df_agg = df_1.loc[:, idx[:, ['labor force(%)', 'Female']]]
df_ag = df_1.loc[:, idx[:, ['labor force(%)', 'Male']]]
df_agg
```

```
In [ ]: """creating a sorted Pivot Table of Labor Force and Unemployment to have
an analysis of Countries having Highest and Lowest labour Force and Unemployment rate(Females)"""
```

```
b = pd.DataFrame(df_1.pivot_table(df_agg, index='Year', columns='Country').aggregate(['max','min','median']).max())
b
```

Countries Like Netherlands and Canada are having Highest rate of Female Labour Force and the Lowest being Pakistan and India . The Highest rate of Unemployment among Females is at Spain and Italy(about 50% of total female population). and Lowest is at Pakistan(6.78%)

```
In [ ]: """Craeting another analysis using Pivot table to isllustrate Labor Force and Unemployment among males and to have
an anaysis of Countries having Highest and Lowest labour Force and Unemployment rate(Males)"""
```

```
a = pd.DataFrame(df_1.pivot_table(df_ag, index='Year', columns="Country").aggregate(['max','min','median']).max()
,columns=[' Labour force and Unemployment Male'])
a.sort_values(' Labour force and Unemployment Male',ascending=False)
```

```
In [ ]: """Defined a function to find the ratio between two columns of the dataframe and returned the values as dft,dff,dfm"""
```

```
def umemployment_labor_Ratio(df):
    df_Total = np.round(df['unempoyement(%)','Total'].sum() / float(df['labor force(%)','Total'].sum()),2)
    df_Female = np.round(df['unempoyement(%)','Female'].sum() / float(df['labor force(%)','Female'].sum()),2)
    df_Male = np.round(df['unempoyement(%)','Male'].sum() / float(df['labor force(%)','Male'].sum()),2)
    return df_Total,df_Female,df_Male
```

```
data = pd.DataFrame(df_1.groupby(level = 'Country').apply(umemployment_labor_Ratio),
                    columns = ['unemployment laborforce ratio (Total,Female ,Male)'])
data.sort_values('unemployment laborforce ratio (Total,Female ,Male)',ascending = False)
```

Unemployment ratio to the total labour force in countries are analysed to see that spain and italy are having the highest umemployment rate and pakistan and japan has the lowest

Countries Like Japan and Netherlands are having Highest rate of Male Labour Force and the Lowest being Italy and Mangolia Comparably . The Highest rate of Unemployment among males is at Spain and Italy(about 50% of total male population). and Lowest is at Pakistan(8.16%)

```
In [ ]: """ five countries the highest and lowest level of GDP gross by groupby aggregate function"""
```

```
idx = pd.IndexSlice
df_agg = df_2.loc[:,idx[:,['GDP','gross']]].groupby(level=['Country']).aggregate(['max'])
df_agg_max = df_agg.sort_values([('GDP','gross','max')],ascending=False).head(5)
```

```
print(df_agg_max)
```

ON grouping Countries over last 10 years USA have the highest GDP gross and Mangolia being the least

```
In [ ]: """Slicing the Second Dataframe for different employment Sectors"""

df_emp = df_2.loc[:,idx[:,['Employment','industries']]]
df_agri = df_2.loc[:,idx[:,['Employment','Agriculture']]]
df_Ser = df_2.loc[:,idx[:,['Employment','Services']]]

"""Grouping and applying filter different Employment sectors having
values greater than average Employment value in each Sectors"""

a = df_emp.groupby(level = ['Country']).filter(lambda x :x['Employment'].mean()>(df_2['Employment','industries'].mean()))
b = df_agri.groupby(level = ['Country']).filter(lambda x :x['Employment'].mean()>(df_2['Employment','Agriculture'].mean()))
c = df_Ser.groupby(level = ['Country']).filter(lambda x :x['Employment'].mean()>(df_2['Employment','Services'].mean()))

display(a,b,c)
```

Highest Employment in Industries are with India and Italy Highest Employment in Agriculture are with India and srilanka Highest Employment in Services are with Canada ,Spain and Britan

```
In [ ]: """Total average change in the GDP and Emplyement from 2014"""

df_average = (df_2.loc[idx[:,['YR2014','YR2016','YR2018'],:],:])
pd.DataFrame(df_average.groupby(level = ['Country']).transform(lambda x: x - x.mean()).sum(),columns = ['Values'])
```

AN overall average deviation of attributes over last 6 years shows that the advancements in In service and Industries are pretty high.

```
In [ ]: """Analysing the combined dataframes
Merging the two dataframes df_1 and df_2 to create df_new"""

df_new = df_1.merge(df_2,on = ['Country','Year'],how = 'outer')
df_new
```

```
In [ ]: """Grouping the two dataframes with unemployment and GDP factors and
```

```
finding the overall correlation between these two factors"""
```

```
df_new.groupby(level = 'Country')[['unemployment(%)', 'Total'], ['GDP', 'gross']].mean().corr()
```

It can be estimated that Unemployment can influence the GDP of any nation Negatively by 11.11%

```
In [ ]: """Mapping the merged dataframe with the external Dictionary 'country Groups ' and performing Groupby to have an
        overall analysis about unemployment ,labour Force and GDP in different regions of the world
        """
```

```
df_new['Region'] = df_new.index.get_level_values(0).map(country_groups)
df_new.groupby('Region')[['unemployment(%)', 'Total'], ['GDP', 'gross'], ['labor force(%)', 'Total']].sum()
```

It can be inferred that European Countries have the highest rate of Unemployment and Labour Force. North American Countries have the lowest Rate of Unemployment and Highest GDP. Whereas Asian Countries maintain a moderate Unemployment rate, Labour Force and GDP.

```
In [ ]: df_emp = df_new.groupby('Region').apply(lambda x : x.mean())[['Employment', 'labor force(%)']]
```

```
In [ ]: df_emp
```

On Analysis North American Countries are having the highest Labour Force 75.27%. To enhance their GDP gross and PPP highly they focus and increase the employment in service sectors rather than the conventional agriculture and industrial sectors of work. South Asian Countries being least in the list give more importance to the agricultural and industrial sector rather than Services.

END