Gender Dataset Pre-processing Steps

The raw data was downloaded from the World Bank Gender Statistics Portal. The portal allows users to select the time period, indicator, and countries for which the data is required. We selected data for all indicators and all countries for the period 2012-2020.

This document gives the details of the steps used to process the raw data downloaded from the portal to create the final dataset.

Please note, students **do not need** to understand the data pre-processing steps to understand the case study. This document is for internal use only for anyone who is curious to know the source of data.

Data pre-processing steps:

1. The raw dataset was downloaded from the portal for the period 2012-2020

2. The data was downloaded for each year for all indicators in separate excel files for the period 2012-2020 and then the files were merged to create a single file.

```
Sample Python code:
df = pd.DataFrame()
for file in files:
       if file.endswith('.xlsx'):
       df = df.append(pd.read_excel(file), ignore_index=True)
```

3. The list of all indicators was extracted

```
Sample Python code:
df['Series Name'].unique()
```

4. The indicators with missing data or data values that had “....” instead of the actual value were removed.

```
Sample Python code:
df = df[(df['Value'] != '..') & (~df['Value'].isnull())]
```

5. Out of the remaining indicators, the important ones were shortlisted and grouped together under various categories. The following categories were created: Education,

Health, Finance, Labor, Access to technology, Demographics, Index, and Other. Following this, each indicator was assigned to a category.

```
Sample Python code:

education=[ 'Children out of school, primary, female',
 'Children out of school, primary, male',
  'Expected years of schooling, female',
 'Expected years of schooling, male','Literacy rate, adult female (%
of females ages 15 and above)',
 'Literacy rate, adult male (% of males ages 15 and above)','Rate of
out-of-school youth of upper secondary school age, female (%)',
'Rate of out-of-school youth of upper secondary school age, male
(%)','School enrollment, primary, female (% gross)',
  'School enrollment, primary, male (% gross)',
  'School enrollment, secondary, female (% gross)',
   'School enrollment, secondary, male (% gross)' ]
```

A similar python code was used to assign all indicators to a specific category.

```
Sample Python code:
df['Category']=np.where(df['Series Name'].isin(health), 'Health',
          np.where(df['Series Name'].isin(education), 'Education',
          np.where(df['Series Name'].isin(finance), 'Finance',
          np.where(df['Series Name'].isin(labor), 'Labor',
          np.where(df['Series Name'].isin(demographics),
 'Demographics', np.where(df['Series Name'].isin(tech), 'Access to
   technology', np.where(df['Series Name'].isin(index), 'Index',
                             'Others'
                   )))))))
```

6. The columns that were not necessary for analysis were dropped. For example, the columns, series code and country code were dropped from the raw data.

```
Sample Python code
df=df.drop(['Series Code','Country Code'],axis=1)
```

7. The remaining columns were renamed and the datatype for values was changed to integer from string for the purpose of calculation.

```
Sample Python code:
df.rename(columns = {'Series Name' : 'indicator', 'Country Name' :
```

```
'country', 'Value' : 'value', 'Year' : 'year', 'Category' :
'category' }, inplace=True)
```

```
df['Value']=pd.to_numeric(df1['Value'], errors='ignore')

df.to_csv(‘gender_data.csv’)
```

8. The final dataset was converted to a CSV and then uploaded on our partner platform Mode to create the data table - career_nub.gender_data