# WEB SCRAPING WITH KNIME

## Introduction :

web scraping is the most important thing in Data science. our objective in this article is to make a simple workflow in knime scraps the data from the website. Here we have to take a live website of corona cases and other information for Indian you can find the website here.

## x path :-

* so our next step is to parse the xml file to extract the table data from it x-path node helps you out to parse the xml file, as you can see below the configuration of x path we have to specify the expath

* using the add path button you can add a different path and xpath summary displays how many path you here

Pivoting and ungroup :-

Now our next step is to transform the table into a data file we use pivoting node. and after this, we have to ungroup the column using ungroup node

web page retriever node :-

* web page Retriever node provides the facility to content with the website and generate the XML file, so as you can see in in the below dialog box of webretriever node you have to just specify your respective URL in the connection setting in our case we have to specify this web site in URL

* you can find a more detailed description of webpage Retriever node from here.

CSV writer :-

our final step is to write a excel (or) CSV writer file to this fetched data and knime provides facilities to do this by simply using CSV writer nodes.

so by executing this full workflow at the end in the CSV or excel you get the fetch data - wise billionaires list

# CORRELATION

## •correlation Analysis :

correlation is a statistical techniaue that can be used to determine if. how stroongly pairs of variables are associated. correlation is only appropriate for quantifiable data in which numbers are meaningful such as companies or ordinal data. It cannot be used for purely categorical data for which we have use conrigency table analysis If one variable ivedes from its mean does the other variable viate from its mean in does the other opposite reaction. this can be assembled by measuring ariance, houlevet, this not standerdiosed. we measure the covariance of two variables matter, if we convent the centimetery, we get some relationship but which a completely ferent convariance value. In order to oversome , standardised covariance is used which is w as pearsons correlation coefficient con"" ranges from -10 to +10 The closer n is t con -1, the mot closely the two iables are related. If n is

close to 0 there is no relationship. If r is (+) then as one variable increases the other also increases. If r is (-) then as one increase. If r is (-F) then as one increased. The other decreases (sometimes referred to as an "increase correlation"

The correlation coefficient (r) should not be confused with $R^2$ coefficient of determination) or R (multiple correlation coefficient as used in regression).

The main assumption in this analysis is that the data have a normal distribution and are linear this analysis will not work well with curvilinear relationships

from now on, we will use the alpha level of .05 because it is most commonly used.