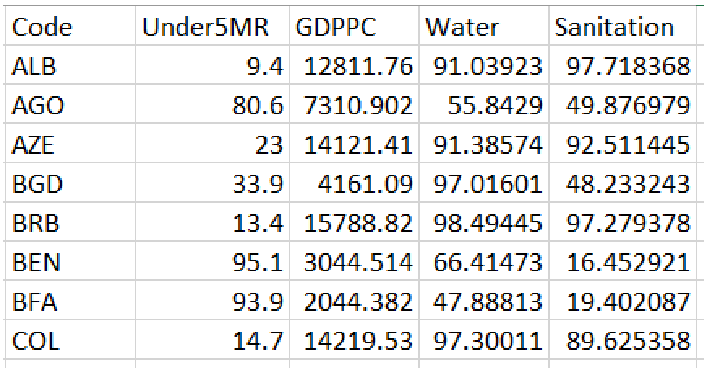
Introductory Econometrics

Tutorial 2

<u>PART A:</u> This section asks you to do some data work . You need to do Part A on your own and answer five questions of the short quiz labelled "Part A Quiz" on Moodle **before** your tutorial in Week 2 **and** attend the tutorial to earn your 1 participation point for Week 2.
You will need to use the VLOOKUP function in Excel. If you do not know what the VLOOKUP function is and what it does, start by watching the two short videos on VLOOKUP on Moodle. This tutorial is about different kinds of data (cross section, time series and panel) that we often work with in business and economics. It is based on the World Development Indicators (WDI) database https://datatopics.worldbank.org/world-development-indicators/, which contains a large number of variables for all countries of the world for every year from 1960 to the present. Of course not all countries have data going back to 1960, and also not all countries collect the same data. So, there are many missing values. WDI has an online databank that allows you to extract and visualise variables that you may be interested in. It also allows you to download the entire data set in bulk in one huge spreadsheet with more than half a million rows.
The entire WDI dataset is a panel dataset because it contains data for a set of countries from 1960 to present. Since the entire data set is huge, I have done some pre-processing and created `WDIextracted.xlsx` that has the following 5 sheets:

- "Basic Water": contains % of population using at least basic drinking water resources for a large number of countries from 2000 to the latest available year.
- "Basic Sanitation": contains % of population using at least basic sanitation services for a large number of countries from 2000 to the latest available year.
- "Under-5 Mortality Rates": contains mortality rate in children under 5 years old (per 1000 live births) for a large number of countries from 2000 to the latest available year.
- "GDP Per Capita": contains GDP per capita for a large number of countries from 2000 to the latest available year. These figures are in US dollars adjusted for purchasing power differences among countries. This is referred to as "purchasing power parity (PPP) dollars". For example, one Swiss franc is equal in value to one US dollar. However, because Switzerland is an expensive country, with one Swiss franc in Switzerland you cannot buy as much as you can buy with one US dollar in the USA. A 1 Swiss franc in Switzerland may only have the purchasing power equivalent to 90 US cents in USA. These considerations have been taken into account and adjusted for, so the dollar values have the same purchasing power and hence are directly comparable across countries.
- "A sample of countries": A list of 26 countries chosen randomly from countries with *less than the median GDP per capita in the world in 2017* (less than 50 dollars per person per day).

The purpose of this pre-tutorial exercise is for you to learn to use the VLOOKUP function of Excel to make a *tidy data set* for a group of countries from your messy WDI data set. A *tidy data set* is a data set that has the variables that we want to use in our data analysis in columns, and each row of it contains one observation on those variables. The following figure shows a snap shot of the top 8 rows of a tidy data set containing under-5 mortality rate, GDP per capita, % population with access to basic drinking water and % population with access to basic sanitation services for a cross section of countries in 2017.



Code	Under5MR	GDPPC	Water	Sanitation
ALB	9.4	12811.76	91.03923	97.718368
AGO	80.6	7310.902	55.8429	49.876979
AZE	23	14121.41	91.38574	92.511445
BGD	33.9	4161.09	97.01601	48.233243
BRB	13.4	15788.82	98.49445	97.279378
BEN	95.1	3044.514	66.41473	16.452921
BFA	93.9	2044.382	47.88813	19.402087
COL	14.7	14219.53	97.30011	89.625358

The first row contains variable names. The first column often contains an observation identifier. For cross section data an identifier is some unique tag for that observation. In the above figure, the identifier is three letter country code that the United Nations has assigned to each country. If we were dealing with data for students in a class for example, student ID would be a good identifier. While these identifiers are not used in the analysis, they are useful for merging data from different sources or for descriptive analysis (e.g. identifying which country has the highest GDP per capita in our sample). For time series data, the first column is usually the date. Fortunately, statistical packages these days are intelligent enough to read dates written in various formats.

Remember that while having data in a tidy format is important for transferring data into a statistical package and analysing it, it is more important to know what our data measure. That is, we need to know what Under5MR, GDPPC, Water and Sanitation stand for, we need to know their unit of measurement, and we need to know the time period that the data correspond to. Also, the country codes are not always easy to guess. So, we need a table of country codes, or we can add a column to our data containing country names, although since some country names have spaces and strange characters in them, they may cause problems when reading the data in some statistical packages (fortunately EViews is intelligent enough to not get confused).

Now the exercise that you need to do:

1. For the countries in the "A sample of countries" sheet, use Excel's VLOOKUP function to get country's under-5 mortality rate, GDP per capita, % population with access to basic drinking water and % population with access to basic sanitation services **in 2017** C to F in front of each country code (actually I have already done under-5 mortality rate for you, so you can use that as a reference). The syntax of VLOOKUP function is:

   = VLOOKUP(lookup_value, table_array, col_index_num, type_of_match)

   The inputs to VLOOKUP are:

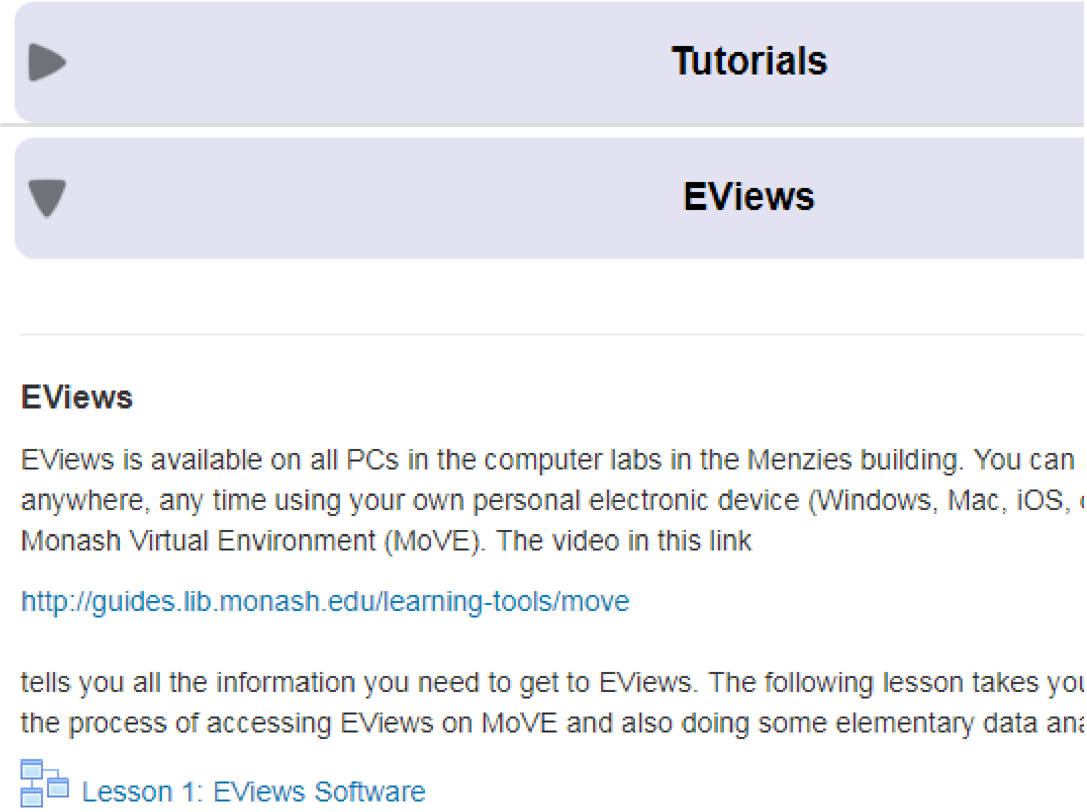
   - lookup_value: the cell that contains the country code (e.g. $B2)
     - table_array: the range of data to search (e.g. 'Under-5 Mortality Rates!$B$2:$X$218)
     - col_index_num: the column that contains the data that you want (relative to the first column in table_array)
     - type_of_match: the accuracy of the match. You always want FALSE here to find exact matches.

   Note the use of $ in the examples. A $ before a column label keeps that column fixed when the formula is copied to other cells. A $ before a row label keeps that row fixed when the formula is copied to other cells.

2. To check if you have done this correctly, unhide a hidden sheet in the `WDIextracted.xlsx` and check the data with what you have created (if you don't know how to unhide a hidden sheet in Excel, google it!). If they don't match, click on different cells and compare the VLOOKUP formulae with yours. If you have any questions, ask on the discussion forum or bring your question to the tutorial.

3. Take the Part A - Tutorial 2 quiz on Moodle. You should submit this quiz before your tutorial and attend your tutorial to get 1 point for participation. It does not matter if some of your answers to quiz questions are incorrect.

END OF PART A. YOU MAY WANT TO REFER TO THE EVIEWS LESSON ON MOODLE FOR A REMINDER ON HOW TO IMPORT DATA INTO EVIEWS.



▶ Tutorials

▼ EViews

EViews

EViews is available on all PCs in the computer labs in the Menzies building. You can anywhere, any time using your own personal electronic device (Windows, Mac, iOS, Monash Virtual Environment (MoVE). The video in this link

http://guides.lib.monash.edu/learning-tools/move

tells you all the information you need to get to EViews. The following lesson takes yo the process of accessing EViews on MoVE and also doing some elementary data an

Lesson 1: EViews Software

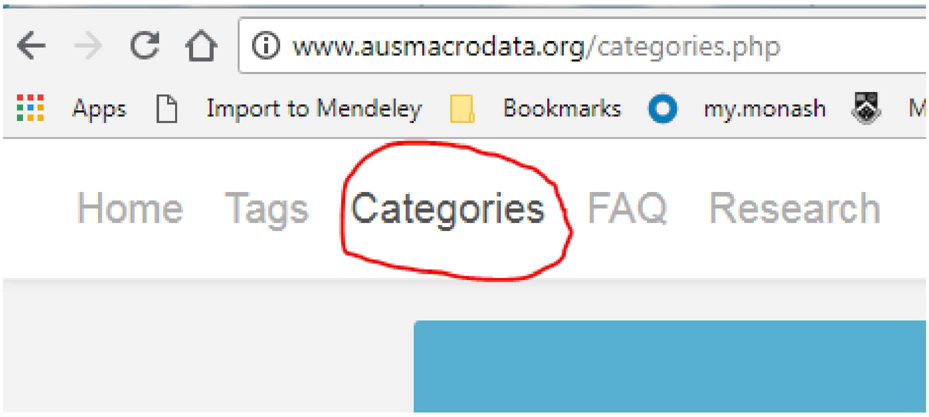
PART B: To be done in the tutorial.

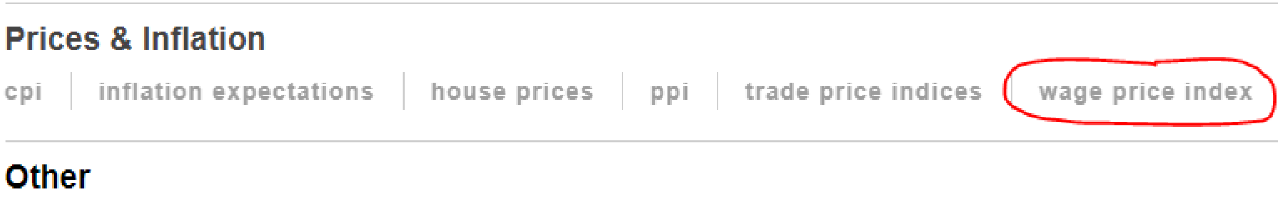
Question 1: *Important topics of the first year statistics: Histograms, scatter plots, and remembering logarithms*: Read the data that you compiled for "A sample of countries" into EViews. When an Excel spreadsheet contains several sheets of data, the first dialogue in EViews allows you to specify which sheet you are interested to read in EViews. If you have not saved your Excel work or had difficulty with it, download "A_sample_of_countries.wf1" from the moodle site and use that.

1. Obtain the summary statistic and histogram of GDPPC. Discuss what you can learn from the histogram and summary statistics (remember that this is not a representative sample of all countries in the world, but is a random sample of countries with GDP per capita of less than 50 dollars a day). If you had a different set of 26 countries with GDP per capita of less than 50 dollars a day, would the summary statistics be the same?

2. We want to explore the association between under_5 mortality and GDP per capita for lower than median income countries. What kind of a graph can give us an insight into the nature of this relationship? Based on this graph, are under_5 mortality and GDP per capita positively or negatively correlated? Is their relationship linear?

Question 2: *Working with time series: plots, trends, seasonality, growth rate (log-returns):* Download hourly wages from www.ausmacrodata.org-> Categories -> wage price index -> the first series that shows up -> Download CSV. Open the CSV file, and tidy up the data set, i.e. only keep the first two columns and delete everything else, and give a better name for the second column, such as WAGE. Save the CSV file and read it in EViews. Note that EViews automatically realises that you have quarterly data.





1. Plot the WAGE series (plotting a time series means producing a line plot of the series in which the x-axis is time. In EViews, clicking on a series opens a window that shows the value of the series in a spreadsheet. This window has a menu bar. Under View is Graph. And the default for Graph is a line plot). What can we learn from this plot? (Note that this is a wage index with average hourly wage in 2008-2009 financial year normalised to 100. So, the numbers are proportional to hourly wage, but are not the dollar values).

2. If we were interested in forecasting hourly wage index in the next period, would sample average be a good forecast? Suggest more appropriate forecasts.

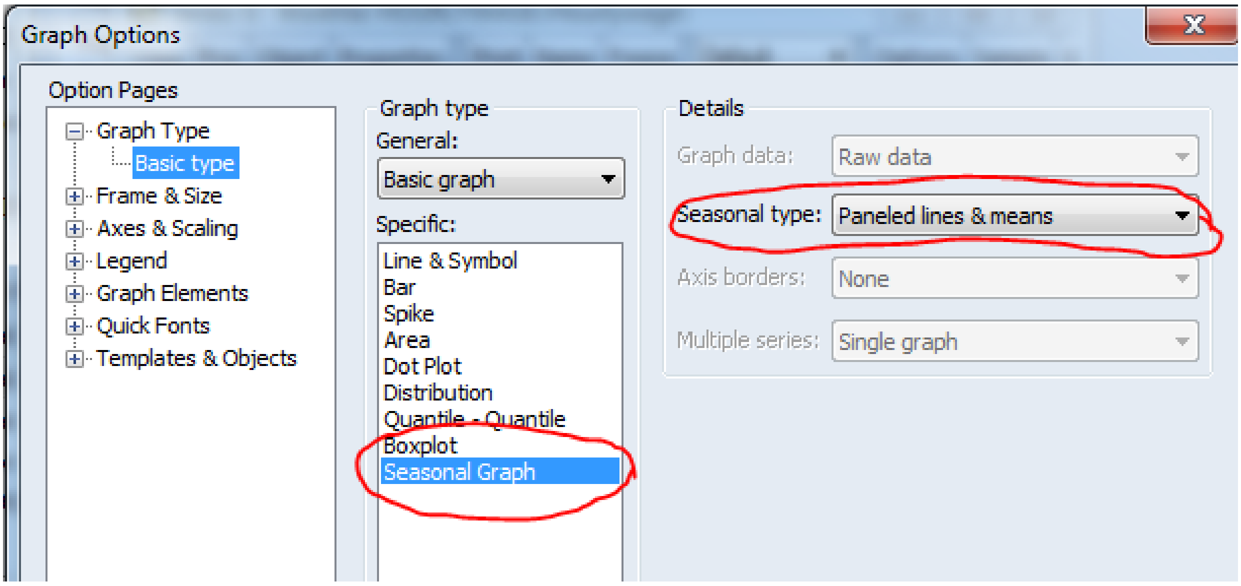
3. In most financial or economic time series, trend is so dominant that it is the only thing that we can immediately see and all other aspects of the time series are dwarfed by its trend. To see other aspects of the series we have to remove its trend. One way is to make a model with time as an explanatory variable (we will do this later in the course). Another way is to compute the growth rate, in this case $g_t = 100 \times \frac{wage_t - wage_{t-1}}{wage_{t-1}}$ (multiplication by 100 is just to express it in percentage points). A more prevalent way of calculating the growth rate, in particular in finance, is to use what is known as "log-returns"

$$g_t = 100 \times \Delta \log(wage_t) = 100 \times (\log(wage_t) - \log(wage_{t-1}))$$

These two methods of calculating the growth rate produce values that are close to each other as long as growth rate is less than 10% in absolute value. EViews has a built in function 'dlog(X)' that computes the difference of logarithm of X. Generate the growth rate of hourly wage using the log-returns formula. Open this series. Why is the first value of this series NA?

4. Plot the growth rate of wage. What does this plot tell you?

5. Look at the seasonal plots of the growth rate of wage. To produce seasonal plots in EViews, in the series window, View -> Graph, and choose Seasonal Graph (the last option under Graph type). There are two seasonal plots: one that plots each season in a different panel side by side and shows average growth rate for each season. Another type shows four line plots, one for each season, overlaid on one graph. Look at both plots and discuss what you learn from these plots.



Graph Options

General: Basic graph

Specific: Seasonal Graph

Graph data: Raw data

Seasonal type: Paneled lines & means

Axis borders: None

Multiple series: Single graph

6. Look at the time series plot of the growth rate of wage again. Mentally adjust for seasonal variation. Do you see that the wage growth has been declining since 2010? Should that by itself worry us? What else do we need if we are worried about the value of one hour of work?