MA26620: Practical 4

Electricity Generation Data and Assignment One

1 Electricity Generation Data

1.1 The Data

The dataset 'elecgen' that is on the module webpages (under Practical worksheets) is a modified subset of data collated by the scientific online publication Our World in Data[1]. The data concerns electricity generation for different countries and territories between 2005 and 2015 inclusive. Below is a description of fields in the dataset:

iso	Three-letter ISO 3166-1 alpha-3 country code
country	Country or territory name
year	Year
region	Region of the world where country is situated
subregion	Subregion where country is situated
population	Population of the country
group	Income group of country, as assigned by the World Bank (1=low, 2=lower-middle, 3=upper-middle, 4=high)
gdp	Gross domestic product of country (in billions of dollars)
total	Total annual generation of electricity (in terawatt-hours, TWh)
coal	Annual electricity generation from coal (TWh)
oil	Annual electricity generation from oil (TWh)
gas	Annual electricity generation from gas (TWh)
hydro	Annual electricity generation from hydroelectricity (TWh)
nuclear	Annual electricity generation from nuclear power (TWh)
solar	Annual electricity generation from solar power (TWh)
wind	Annual electricity generation from wind power (TWh)
pc	Annual electricity generation per capita (i.e. per head of population) (kWh/person)
ren	Percentage of electricity generated from renewable sources (%)

The dataframe also contains variables named `coal_percent`, `oil_percent`, `gas_percent`, `hydro_percent`, `nuclear_percent`, `solar_percent`, and `wind_percent`, which give the percentage of annual electricity generation from each source. Renewable generation also includes sources not listed separately here, so `ren` need not equal the sum of the hydro, solar and wind percentages.

The income groups are the classifications supplied with this dataset and are constant for each country throughout the period; do not assume they describe changes in income classification from year to year. Electricity generation is not the same as consumption: countries may import or export electricity, and losses also occur before electricity reaches its users.

1.2 Preliminaries

Begin by importing the data as a dataframe called `elecgen` and attach it using the usual method. Be sure to indicate that the first row of data gives column headings by selecting 'Yes' for the 'Heading' option, otherwise you'll end up with meaningless variable names like `V1`, `V2`, `V3` etc.

[1] Hannah Ritchie, Max Roser and Pablo Rosado (2022), "Energy". Published online at OurWorldInData.org. Retrieved from `https://ourworldindata.org/energy`.

We should first get a quick overview of the dataset so we know what kind of data we're dealing with. Click on `elecgen` in the top-right pane; the data will appear in RStudio's top-left pane. Have a quick scroll down and across the data. It's a good idea to run the command `summary(elecgen)` too, which will show you summary statistics for each field. This is a good habit to get into; for one thing, eyeballing the summary might sometimes reveal an error in the data. For instance, if you saw a year of '3013', you'd know that the data inputter had probably made a typo. This dataset contains missing values, and rounding can cause small inconsistencies between generation totals and percentages. Some zero generation values may also reflect unavailable data, so be cautious about interpreting them as genuine zeros.

1.3 A sensible way of working

As in previous practicals, click File > New File > R Script. The upper left panel becomes a notepad where you can type commands. So in today's practical and while you're working on the assignment, type your commands in here rather than in the Console pane. When you've typed a command, clicking the 'Run' button will run everything that's on the currently selected line. Equivalently, you can press Ctrl+Enter to run the command on the currently selected line. If you want to run more than one command at once, highlight several commands and then click 'Run' or press Ctrl+Enter.

When you make a typo or your command isn't quite perfect, edit it in the notepad and then run it again.

To reiterate the advantage of working using this notepad rather than just typing things into the Console window: well, you can save your notepad at the end of a session (just click the Save icon in the notepad pane—it's the one immediately above the first line of your notepad). When you do, **be sure to save it somewhere you can access again, such as your university OneDrive or M: drive if available**. Then, the next time you work on your assignment, all you'll need to do is load the data (if it isn't already there), enable ggplot2 in the Packages tab (a step which isn't even necessary if you put `library(ggplot2)` at the start of your notepad), load the notepad that's full of your lovely, polished commands, highlight everything and click run, and R will remake every subset and produce every plot that you've made.

1.4 Tables and/or barcharts

Let's begin to analyse the data by using the tables and barcharts methods that we have learned in previous practicals.

The full dataset contains 11 rows for each country, one for each year between 2005 and 2015 inclusive. This will be useful if we wish to investigate how electricity generation has changed over time. To keep things simple to start with though, let's limit ourselves to just the most recent year in this dataset: 2015.

To do this, let's make a subset of the data like we did in Practical 2. Run the following to create a new dataframe consisting solely of the 2015 data (let's call this new subset `e15`):

```
e15 <- subset(elecgen, year == 2015)
```

Note the double equals-sign; this is R's syntax (and many other languages) for an "is equal to" test, which does not use a single equals-sign since that can be used to denote assignment of a variable (e.g. `a=42`, even though in R we would more commonly use `a<-42`).

Detach `elecgen` and attach `e15`. Now try producing a frequency table showing the number of countries in each income group, using `table(group, useNA = "ifany")`; see the next page for a discussion of how to deal with missing values in the data. You could also display this information as a barplot as we saw in Practical 2, making sure that the income groups are clearly labelled. Consider which presentation is more informative, and whether the missing values should be included.

1.5 Boxplots

Let's use the ggplot2 package that we used last time to create a boxplot with separate boxes for each region. (Note: you will first have to tick "ggplot2" in the "Packages" tab, or run `library(ggplot2)`. If it doesn't appear in the Packages list, repeat the installation instructions from the previous practical.)

To get you started, a basic boxplot showing the percentage of electricity generation coming from renewable sources by region can be made using the command:

```
ggplot(e15,aes(x=region,y=ren))+geom_boxplot()
```

Edit the plot nicely (adding suitable labels, title, and so on), copy and paste it into Word, and add some comments on what can be gleaned from the plot.

How about a boxplot showing the percentage of electricity generation coming from renewable sources, split by income group (you'll have to use `factor(group)` in your command so that R knows that `group` is a categorical variable, as we did in Practical 3). Pick out some key features of what the boxplot shows in your comments on this plot. For instance, which groups appear to differ in median or spread, and how much do their distributions overlap? Are the distributions fairly symmetrical, or skewed? A boxplot alone does not establish statistical significance.

Remember that these plots give each country equal weight. The median of the country percentages is not the percentage of all electricity in the region that comes from renewable sources.

Make any other boxplots that you think may be interesting (you're spoilt for choice). Remember to be selective regarding which of these you include in your final report.

A note on dealing with missing values in R

Real datasets often contain missing values, representing information which is unavailable or has not been recorded. In R, missing values are represented by `NA` (short for "not available"). It is important to distinguish missing values from zeros: a country for which no electricity generation data is available is not necessarily a country which generates no electricity. We can identify missing values using the command `is.na()`. For example, `sum(is.na(e15$gdp))` tells us that 47 of the 213 countries and territories in `e15` have missing GDP values. The command `summary(e15$gdp)` also provides useful information about the distribution of GDP, including the number of missing observations.

Missing values can cause problems when performing calculations. For example, `mean(e15$gdp)` returns `NA`, since R cannot calculate the mean of a set of numbers containing missing values. We can instead use `mean(e15$gdp, na.rm = TRUE)`, where the argument `na.rm = TRUE` instructs R to remove missing values before performing the calculation. Missing values can also cause problems when constructing frequency tables. By default, `table(e15$group)` excludes missing observations, so the frequencies add up to only 204 rather than 213. To include missing values as a separate category, use `table(e15$group, useNA = "ifany")`.

When producing graphs using ggplot2, observations with missing values in variables required for the graph are generally omitted automatically, sometimes accompanied by a warning. For example, `ggplot(e15, aes(x = region, y = gdp)) + geom_boxplot()` produces boxplots of GDP for the different regions, but excludes countries with missing GDP values. This means that the boxplots are not necessarily based on all the countries in each region. We can investigate this using `table(e15$region, is.na(e15$gdp))`, which shows how many GDP values are available or missing in each region. In particular, Oceania contains 17 countries and territories, but GDP data is available for only two of them!

It is therefore important not to remove missing observations without considering the consequences. If only a small number of observations are missing, excluding them may have little effect on the results. However, if missing values are concentrated in particular groups, the remaining observations may not be representative of the original dataset. In the GDP example, comparisons between regions should be treated cautiously because the proportions of missing observations differ considerably. Missing values should not normally be replaced with zero, and any conclusions should take account of how much data is missing and whether the missing observations could affect the analysis.

1.6 Scatterplots

Let's try to make a scatterplot showing the relationship between a country's GDP and its per-capita electricity generation:

```
ggplot(e15,aes(x=gdp,y=pc))+geom_point()
```

You'll see that most of the points are nestled in the bottom left corner. This is because most of the world's countries have fairly small GDP and small per-capita electricity generation compared with the largest values in the dataset. As we saw in a previous practical, if one variable on a scatterplot (or both) is strongly positively skewed, visualisation is often improved by taking logarithms of one (or both) variables.

Logarithms require positive values. Let's therefore make a subset containing only rows with positive, non-missing values of both quantities:

```
e15log <- subset(e15, !is.na(gdp) & !is.na(pc) & gdp > 0 & pc > 0)
```

Here, the "!" means "not", so `!is.na(gdp)` means "`gdp` is not a missing value". We can now make the plot:

```
ggplot(e15log,aes(x=log(gdp),y=log(pc)))+geom_point()
```

Although the linear relationship between these log-transformed variables does not seem to be particularly strong, we can see from this plot that there's a tendency for countries with higher GDP to have higher per-capita electricity generation. Of course we could calculate the correlation coefficient, R, between these two log-transformed quantities to quantify this:

```
with(e15log, cor(log(pc),log(gdp),use="complete.obs"))
```

The `with(e15log, ...)` part tells R to use the variables in `e15log` for this calculation, even though `e15` is still attached; you might find this useful if you end up having lots of different subsets in play. The `use="complete.obs"` option tells R to use only rows for which neither quantity is missing; here we have already removed those rows when making `e15log`.

How about using some colour to add some more information to your plot? Perhaps colour the points by which region each country is in.

As a further investigation, does total GDP measure how wealthy the average person is? You could calculate GDP per person by dividing GDP by population, remembering that GDP is recorded in billions of dollars, and investigate whether this gives a clearer relationship with electricity generation per person. Remember that association does not establish causation.

1.7 Investigating development over time

Recall that the original dataset `elecgen` contained data from 2005–2015 and yet so far we've only looked at 2015 data. Let's now see whether we can investigate how things have developed over time.

How about looking at the percentage of electricity generated by wind in Northern European countries? To investigate this, let's first make a subset of countries in Northern Europe:

```
neur <- subset(elecgen,subregion=="Northern Europe")
```

Then let's make a boxplot:

```
ggplot(neur,aes(x=factor(year),y=wind_percent))+geom_boxplot()
```

How does this compare with Sub-Saharan Africa, say? Use the same vertical scale if you want to compare the magnitudes directly. If most observations are zero, a box may collapse to a line; that is a feature of the data, not necessarily an error in your command.

If we're making many such comparisons, we're going to have to make lots of subsets which might soon become tedious. An alternative is therefore to just specify the subset within part of the ggplot command as follows:

```
ggplot(elecgen,aes(x=factor(year),y=wind_percent))+
    geom_boxplot(data=subset(elecgen,
        subregion=="Sub-Saharan Africa"))
```

If you're comparing a fairly small number of countries, typically around six or fewer, a line plot can also be a good idea. For instance, let's consider countries with a long coastline on the northern Mediterranean: Spain, France, Italy, Croatia, Albania, Greece:

```
nmed <- subset(elecgen, iso %in% c("ESP","FRA","ITA","HRV","ALB","GRC"))
```

Then we make the plot, which uses a combination of a line and points:

```
ggplot(nmed,aes(x=factor(year),y=wind_percent,group=country))+
    geom_line(aes(colour=country),linewidth=1)+
    geom_point(aes(colour=country,shape=country),size=3)
```

When commenting on changes, distinguish a change in percentage points from a percentage increase. Also remember that these are shares of electricity generation: a change in a share need not imply the same change in the amount generated. Successive years for the same country are related observations, rather than independent samples of new countries each year.

1.8 Further investigations

You're now free to investigate the data further in whatever way you wish. The plots we've made so far are a good start, but you should now investigate some other aspect of the data of your choosing. The `subset` command will doubtless be useful, as we've seen many times above. If you can't immediately see how to work out how to do something, do make good use of search engines; there are many R and ggplot guides online and chances are you can find the answer through some careful Googling/Binging/DuckDuckGo-ing or similar. Previous practical worksheets should also be a valuable resource.

2 MA26620 Assignment 1

In this module, the exam is worth 70% while coursework counts for 30% of the module mark. The coursework will be split into two parts: the report set in this worksheet, worth 80% of the coursework component (i.e. 24% of the overall module mark), and some exercises from a practical in semester two, which will be worth 20% of the coursework component (i.e. 6% of the overall module mark).

Your task is to write a report (in Microsoft Word or similar, not handwritten) on what you have learned about the electricity generation data based on a selection of the material you have gathered in this practical and any further investigations into the data that you conduct. The report should be written for a reader who is a competent statistician but is unfamiliar with the dataset (so for instance, you should refer to income groups as "low" and "lower-middle" etc. rather than "1" and "2" throughout).

The assignment must be handed in via Blackboard as a single PDF file.

Your report must include:

(i) a title for the report;

(ii) an introductory paragraph explaining briefly the nature of the investigation, where the data came from, i.e. the nature of the data and the aims of your analysis;

(iii) the results of your R analysis, in a sensible order (quite probably different from the order you did them in the practical). Explain the purpose of each element and give your comments; use sub-headings where it helps to make the structure of your report clearer;

(iv) a conclusion where you summarise briefly what you have learned and any questions that remain unanswered.

(v) an AI tool-use statement at the very end of your report. A Word document template for this is available on the module webpages.

The emphasis of your analysis should be on what's shown by the data; lengthy political/societal justifications explaining what may have led to various changes in energy generation are not required and are indeed discouraged.

You should include at least one of each of the following: table or barplot (either, not necessarily both), boxplot, scatterplot, line plot. Marks will be awarded for choice, presentation, editing, and commenting on each of these components, as well as the general structure and flow of your project. More precise information about what marks will be awarded for is available on the module website under 'Module Information'. You must also give some thought to the order and structure of the content. Consider whether you have used the same style (e.g. font, type size) when you typed the project, that you have organised the text into suitable paragraphs and that the page breaks occur in natural places. Plots must be made in R/RStudio; plots made with other software do not satisfy the module's R/RStudio-related learning outcomes and will earn no credit.

Consider also whether the plots or statistics you have included do indeed illustrate the points you wish to make or whether there are any others which would do the job better. There may be other questions you wish to follow up about this data and some credit (25% of the overall mark for Assignment One) will be given for any relevant additional investigations. Make sure that all plots are well labelled and well presented. State how missing values have been handled and give the number of observations used where this helps the reader interpret a comparison.

Please note that unnecessary or uninformative graphics will receive no credit and, if excessive, may be penalised. Be selective and always have a reason for your choices. There is no word (or page) limit but in the past, most assignments have been fewer than 10 pages including plots.

While I (Adam) encourage you to work as independently as possible, I am happy to assist with queries received via email before the end of Monday 23rd November. I won't give specific technical assistance beyond that date. This policy is intended to encourage you to not leave working on the assignment until the last minute. Otherwise, it is not fair if someone spends days trying to do something and someone else asks me how to do it the day before the deadline and gets a full answer.

Assignments must be handed in as a single PDF file via Blackboard by 3pm on Thursday 26th November 2026.

2.1 Notes on policies

You should be aware of the university's current Unacceptable Academic Practice (UAP) policy and the guidance on permitted assistance for this assignment. Your submitted work must be substantively your own, and you should not send your work to other students to copy (the copier and copied are considered equally at fault in the UAP policy). You should be able to justify that the work is your own if later questioned on it. Any suspected unfair practice will be dealt with under the university's procedures. The penalties for plagiarism or unacknowledged use of AI are severe, so do not be tempted!

Extension requests will normally not be approved if there is less than 72 hours before the deadline (a special circumstances claim is the usual route in that case). Extensions can be requested either via the formal university extensions procedure or by asking Adam. I recommend the latter in the first instance. You do need a good reason as detailed in the policy; having started working on the assignment too late is not a good reason, so I strongly recommend you aim to get it finished a good amount of time before the deadline.