Introductory Tutorial: Part 1 Describing Data

Introduction

Welcome to this R-Instat introductory tutorial. R-Instat is a free, menu driven statistics software powered by R. It is designed to exploit the power of the R statistical system, while being as easy to use as other traditional point and click statistics packages, through a menu and dialog-based design.

R-Instat is the first product developed under the **African Data Initiative (ADI)**, a collaborative project to support improved statistics and data literacy across Africa and beyond. The overall aim of the African Data Initiative project stretches beyond producing this software, however R-Instat is an important first step in achieving change.

The original target audiences for R-Instat were described in the crowd funding campaign which launched the development. We claimed there was a need for statistics software that is easy to use, free and open source and that encourages good statistical practices.

The "Instat" in "R-Instat" refers to a simple statistics package first developed in the 1980s with similar aims and target audiences as R-Instat, and much of the philosophy of R-Instat is inspired by Instat. Instat included a special menu for the analysis of climatic data and R-Instat follows this tradition, as well as including another special menu for the analysis of public procurement data.

We **strongly recommend** following the installation instructions when installing R-Instat. In this document, we focus on introducing you to using R-Instat once it is installed.

The ADI (R-Instat) Team R-Instat@AfricanMathsInitiative.net

Running R-Instat for the first time

Once you have installed R-Instat, it is time to open it!

When R-Instat runs for the very first time after installation, then if you have not used R before, or an updated version or R has been installed, you may see the message box asking:
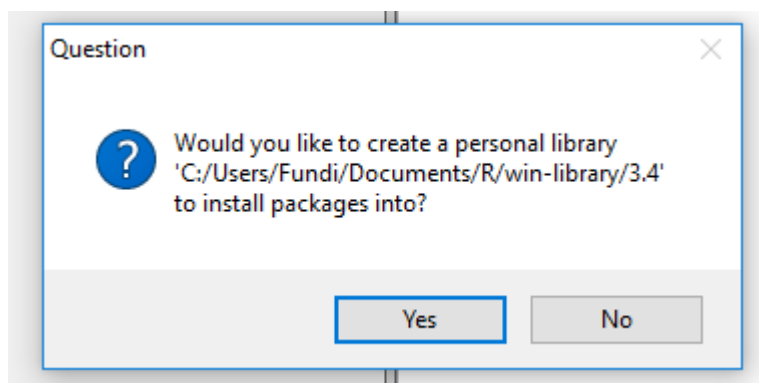
"Would you like to use a personal library instead?".
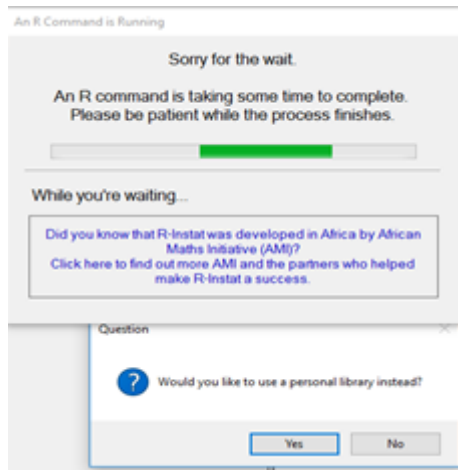
 Click **Yes** for the software to proceed to install the required R packages onto your computer into a folder in your documents.

After clicking **Yes**, another message box may appear saying:

"*Would you like to create a personal library*".

 Click **Yes** again to allow R to install packages to the specified folder.

If you do not see this message box (and no errors appear) then you can assume the R packages were installed correctly as it is likely you already had the necessary folder structures.

When any command takes longer to run (in this case installing packages), you will see the *"Sorry for the wait"* dialog box (above), just to let you know R-Instat is still running and hasn't frozen.

The very first time you run R-Instat, this may take several minutes, as many R packages will be installed. Please be patient! After your first use, this will be much faster as packages will already be installed.

Once the waiting dialog has disappeared you are ready to start using R-Instat!

Exploring R-Instat

This section provides an initial set of examples to help you become familiar with R-Instat and its general features.

1. The Installation.

We hope it went smoothly. Please tell us. Currently R-Instat is a Windows only software. Mac and Linux users could use a virtual Windows machine to install it. We plan to make a cross platform version of R-Instat in the future.

Once installed and opened you should see the screen that looks like this:

Fig. 1: R-Instat main Interface
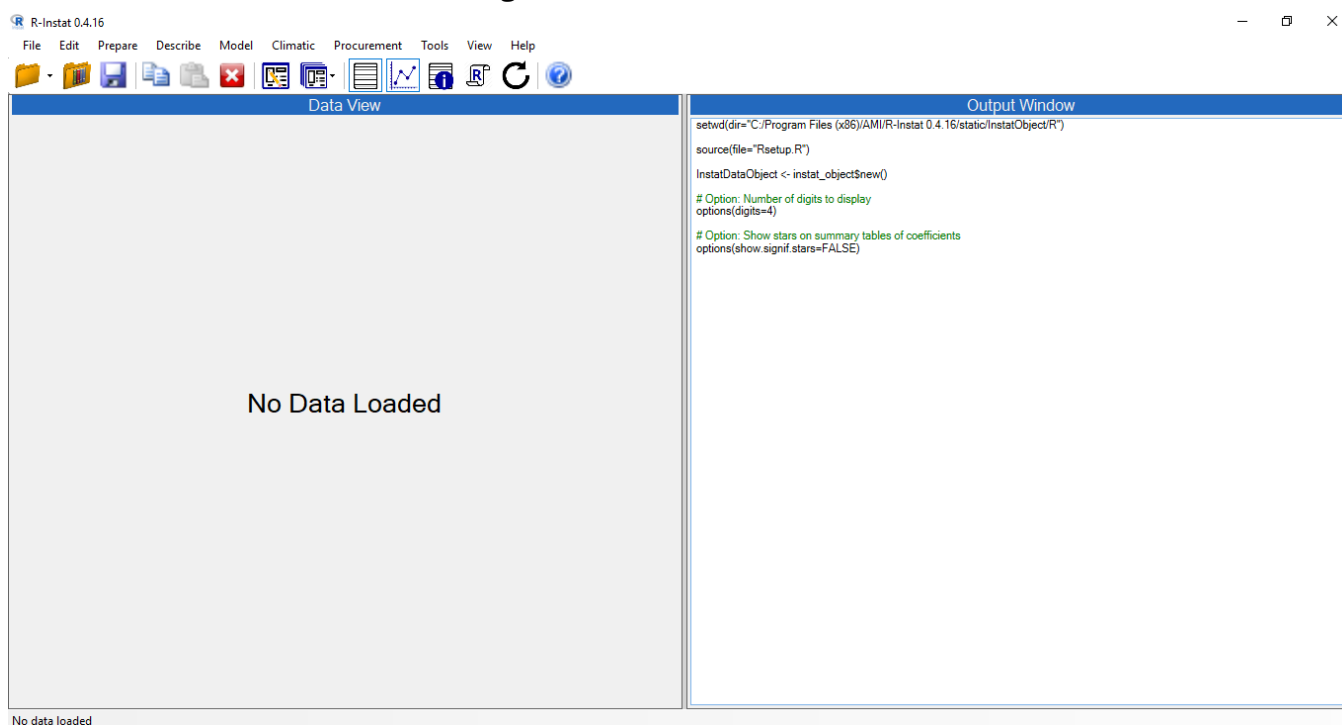
2. A first task - Importing data from the library

☐ Go to *File > Open From Library*.
☐ Click on the *From Package* dropdown and choose *ggplot2*.
☐ Choose the first example, *diamonds* as shown in Fig. 8. You should see that a second *Help* button is now enabled, just below the list of datasets.
☐ Click on that button to get further information about the dataset. This help is shown in a window in a browser. (It is the dataset used by Hadley Wickham, the author of ggplot2, for many of the examples in his own documentation.)

Fig. 2. Using a library dataset
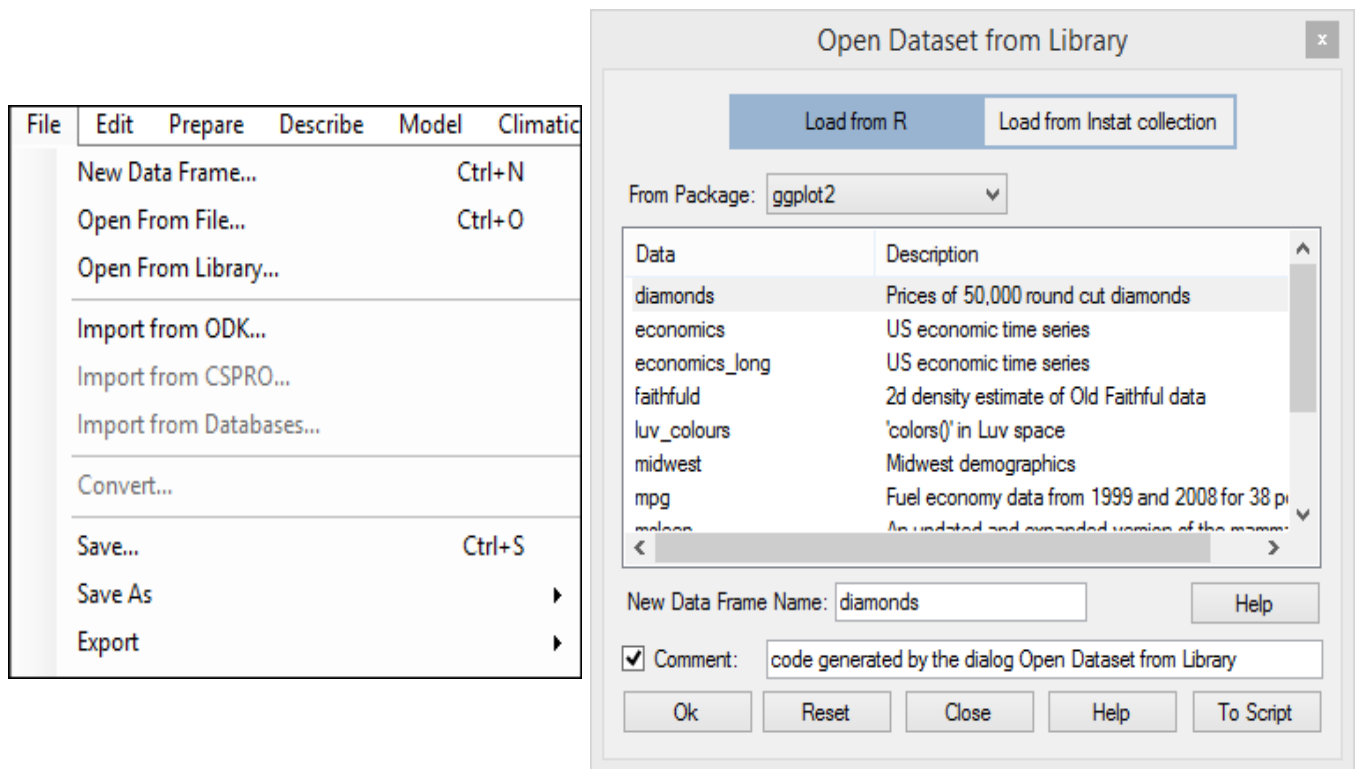
▢ Now return to the dialog, *select the **diamonds** dataset again* and press **OK**.

Fig. 3 The diamonds data

Scroll to the bottom of the data to see it appears to have just 1000 rows. It is just a window onto part of the data frame which is stored in full in R.

 Use **right click** on the bottom tab, Fig. 4.

 Choose the last option, **View Data**. This is one way to see all the rows, also shown in Fig. 4.

Fig. 4. Viewing a data set
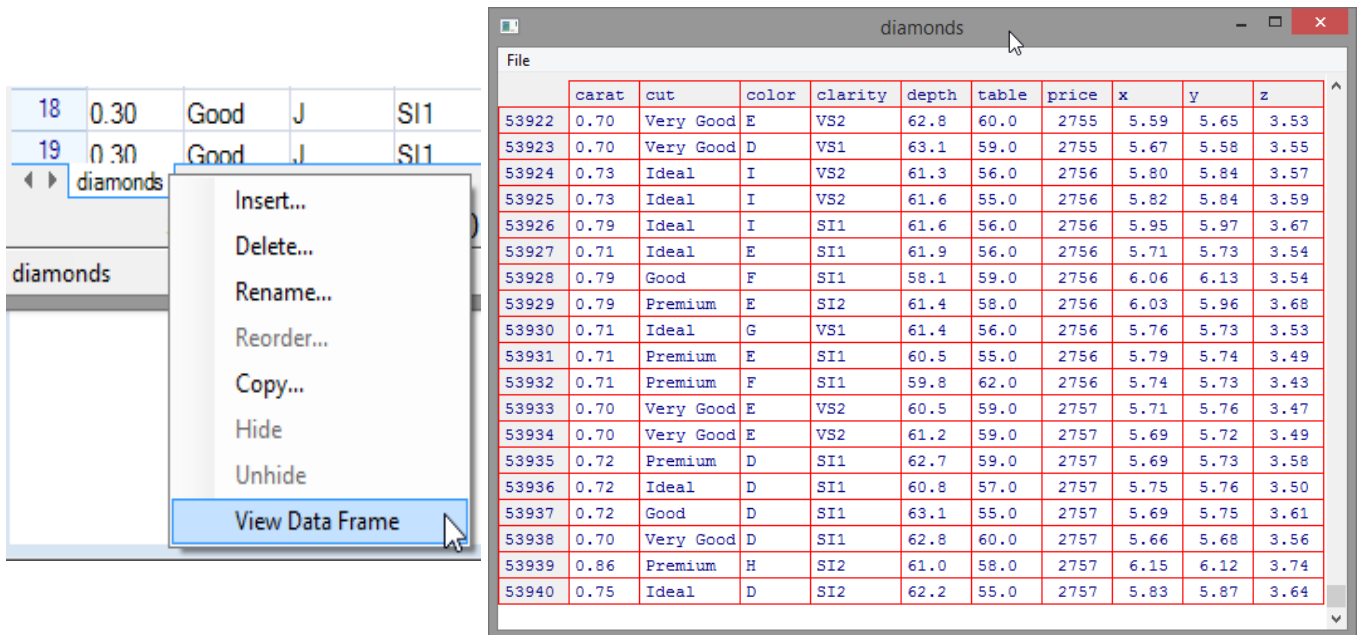
There are 10 columns (variables) of data in this file, of which 7 are **numeric** and 3 are **categorical**. R calls categorical columns **factors** and they are denoted by an "f" after the column name. These categorical columns are actually ordered, for example the second column, namely the **cut** of the diamonds ranges from **Fair** to **Ideal**. Ordered categorical columns are denoted by "**(o.f)**" after the column name in R-Instat.

These data are already well prepared for analysis, so we go straight to R-Instat's Describe menu and show some graphs.

3. Some Graphs

 Go to *****Describe > One Variable > Graph**, Fig. 5.

 * *****Right-click*** in the variables selector and choose the option to **Add All**. (Or you can just select all the columns and then click on ***Add***, Fig. 5.
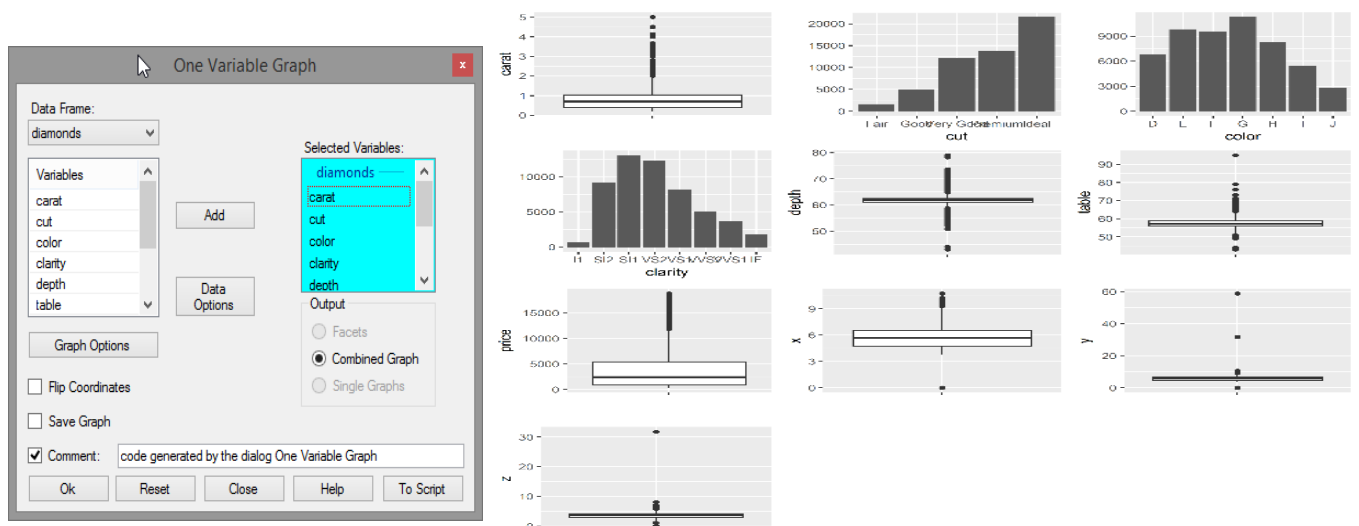
Fig. 5. One Variable graphs dialogue
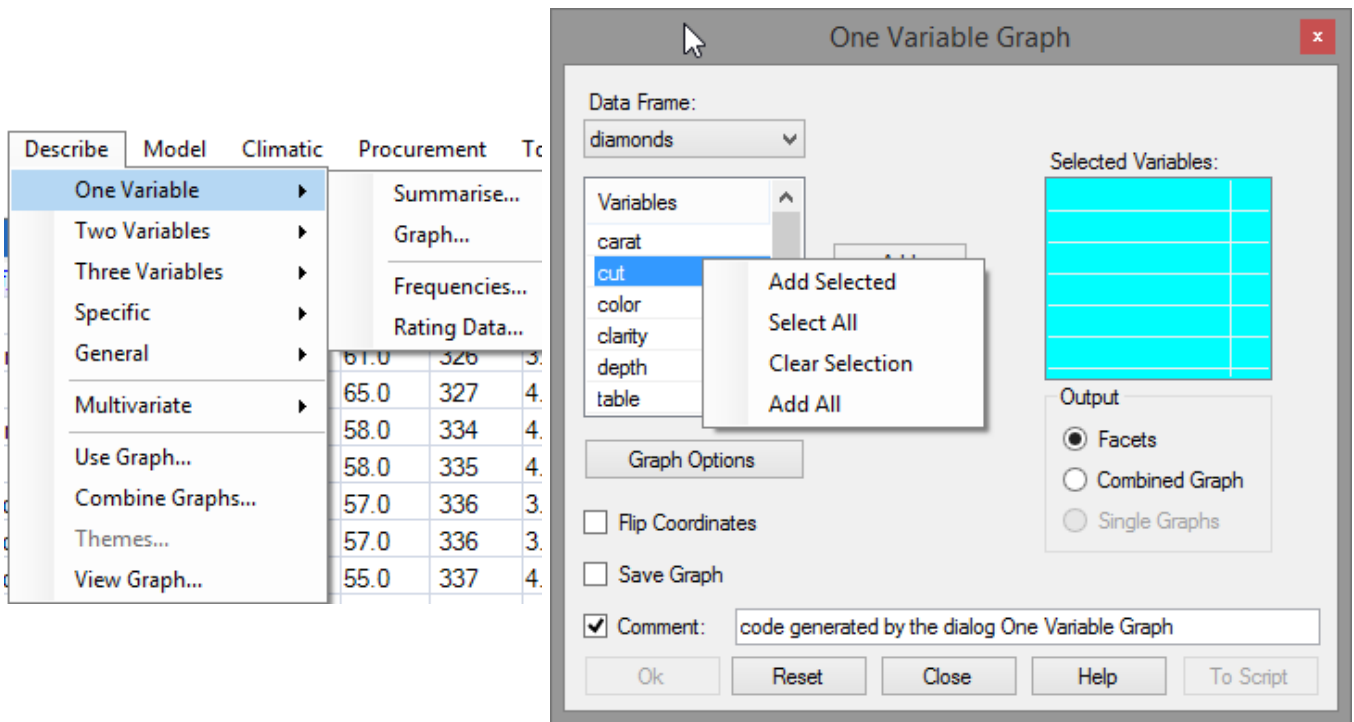
In the dialog in Fig. 5 the radio button changed from *Facets* to *Combine Graph*, see Fig. 6. That is because the selected variables are of different data types. Some columns are categorical while others are numeric.

 Press *OK* to give the results also shown in Fig. 6.

**Fig. 6. One Variable graphs**

You may already be familiar with boxplots? We explain a little about them later, though this tutorial is primarily to show how to use R-Instat, rather than to teach statistics.

Often, the results from using a dialogue can be improved, so you wish to use it again. You could use the same menu options as in Fig. 5, but there is a quicker way.

 Click on the little dialogue picture on the toolbar, see Fig. 7, which takes you back to the previous

dialogue. (Or the next icon lets you return to any of the recently used dialogues.)

Fig. 7. Use the toolbar to return to a dialogue   Or to any of the recent dialogues
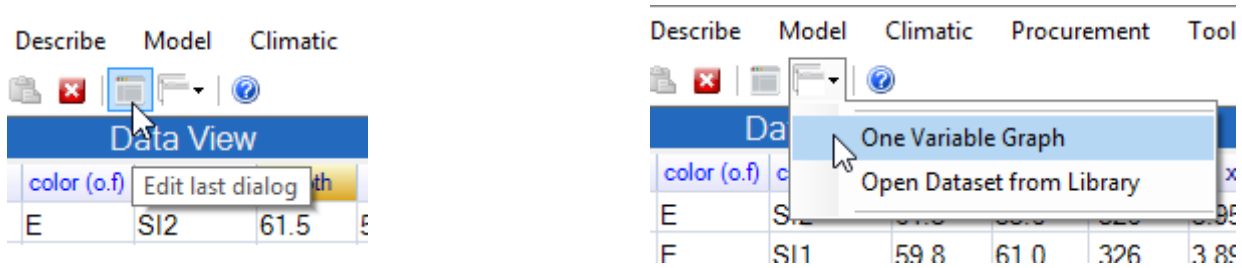
You see the dialogue has "remembered" the settings just as you left it, when you pressed OK. This is often what you want.

 But this time press the **Reset** button at the bottom of the dialogue, to clear all the settings.
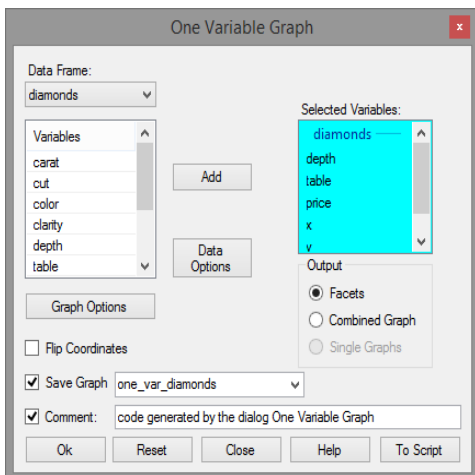 Then omit the first 4 variables and select the last 6, (from to put into the receiver.

As these are all numeric columns the radio buttons on the right have permitted you to do a facetted graph, so you can see what this is!

 Also click on the checkbox to **Save Graph**.
 Name it **one-var diamonds** (Notice you are including a "dash" and a space.)
 Now click **OK**

The dialogue didn't work. Instead it gives a message that "The name cannot contain a space" (or a dash). It is the name of an object in R and these are not allowed.

 Click on **OK** to clear the message box.
 Change the name to **OneVarDiamonds** or perhaps **one_var_diamonds**, Fig. 8, and click **OK** again.

**Fig. 8. The One Variable Graph dialogue
again**   **With a faceted graph**

This shows a **faceted** graph, Fig. 8. This is a multiple graph where the y-axis, by default is the same for all the graphs. This is often what is wanted for a multiple graph, because you don't then need

the axis to be labelled for each variable. However it isn't what we need here. The different variables have very different scales and we need to reflect this in the graph.

 Return to the same dialogue again.
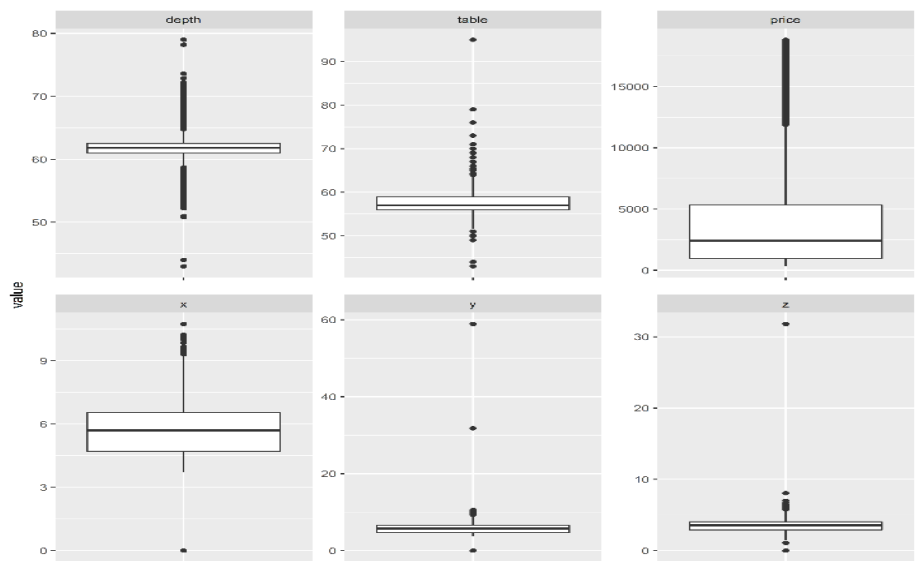 Click on the **Graph Option**s button.

You now see a sub-dialogue with just 2 tabs, Fig. 9. One tab allows you to change the type of graph that is shown.

 Press on the tab labelled **Display** and then click on the **Free Scale Axis**.
 Press on the **Return** button and then on **OK** again, to give the graph also shown in Fig. 9.

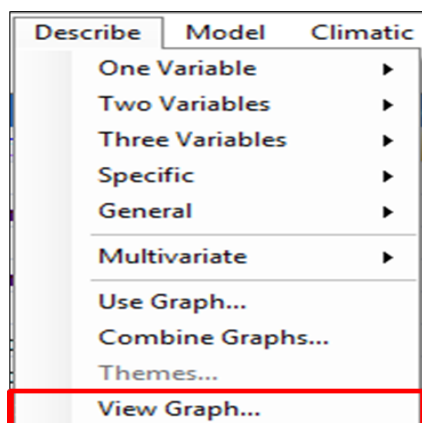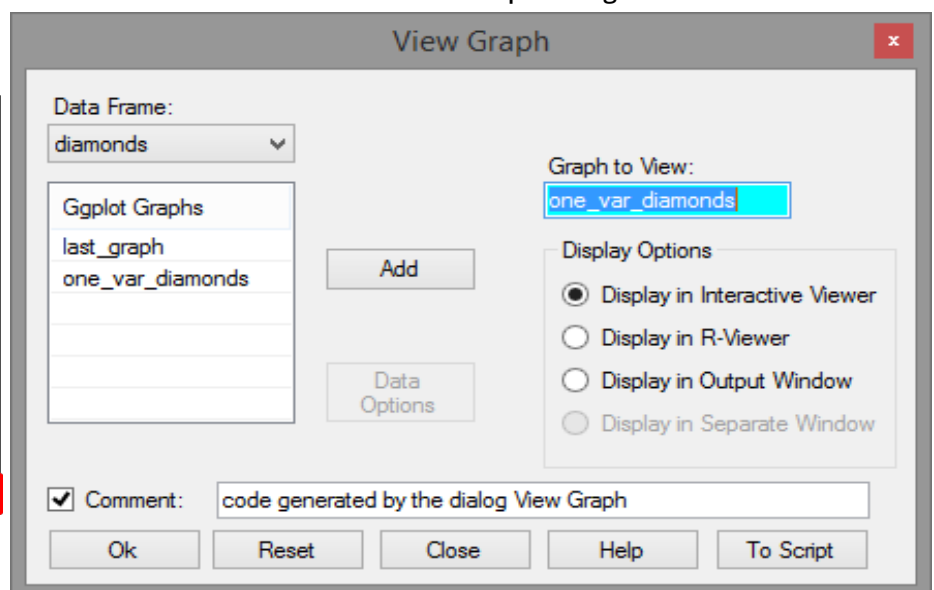**Fig. 9. The One variable graph sub-dialogue**
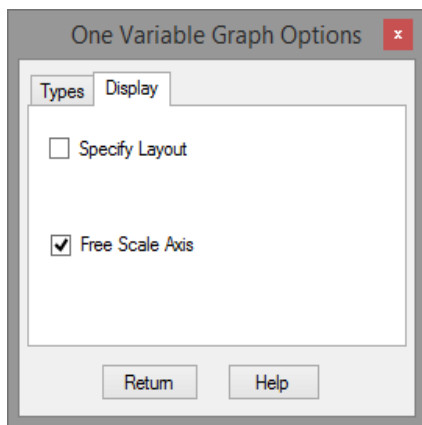
The next graph

 Choose the **Describe > View Graph** dialogue to examine this last graph further, Fig. 10.
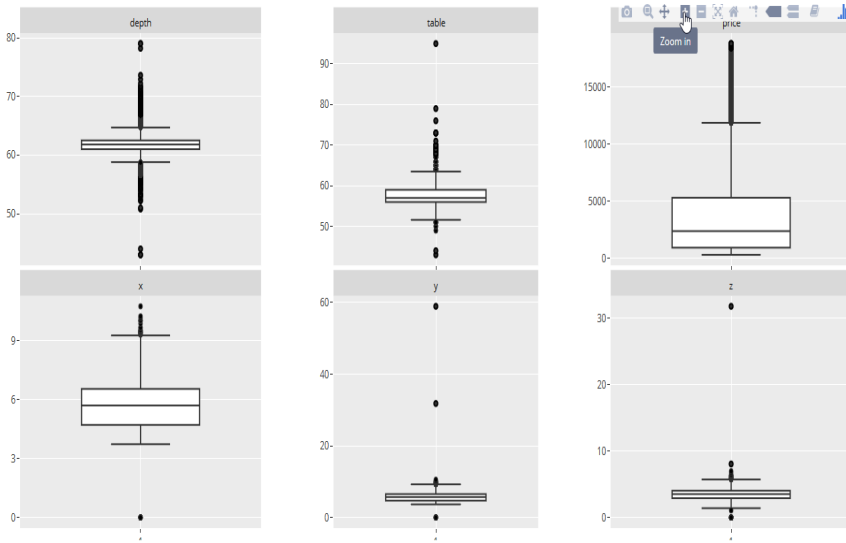
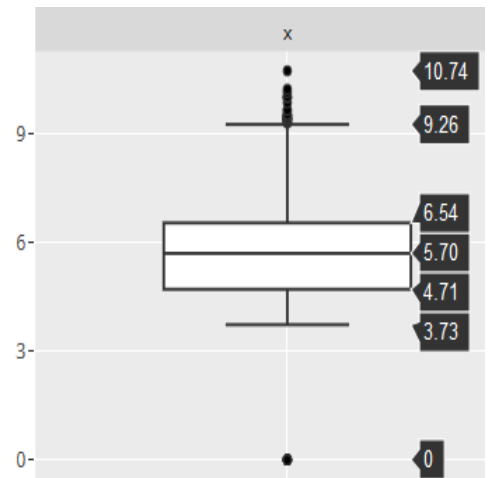**Fig. 10. The Describe menu again**

With the View Graph dialogue

 Press **OK** to show the graphs in a separate (interactive) window, Fig. 11.

Hover over a particular graph to add numerical summaries automatically, Fig. 11.

Fig. 11. The View Graph results   Including a numerical summary
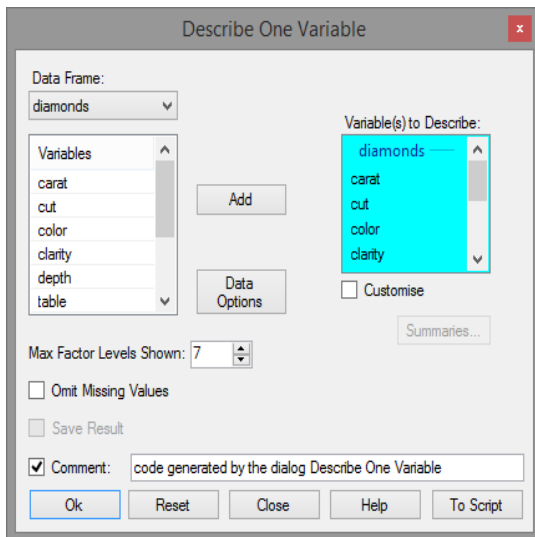

4. Some summaries

Often analyses involve numerical as well as graphical summaries.
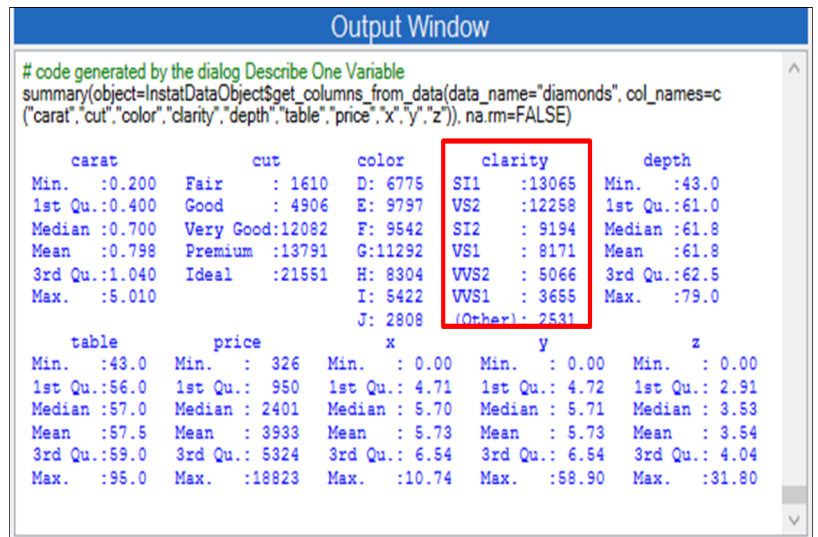
 Go to *Describe > One Variable > Summarise*.
 Select all the variables again (as you did with for the first use of the Graph dialogue), Fig. 12.
 Press **OK** to give the results also shown in Fig. 12.

Fig. 12. The One Variable Summarise dialogue   **With some results**

This is almost right, but the variable marked in a red box in Fig. 12 is not quite clear. It has more than 7 levels (categories), so the remaining ones have been put together.

 Return to the last dialogue.
 In the dialogue, Fig. 12, change the **Maximum Factor Levels Shown** from 7 to 10. Press **OK**.

The levels are now all given for that factor column.

Examine the correspondence between the values given for the x-variable in Fig. 12, with those for the boxplot for x in Fig. 11. They are given together in Fig. 13 to help.

In Fig. 13 the correspondence of the median in the 2 summaries is marked. Are any other values the same? Is the correspondence useful to understand (or to teach) what a boxplot provides.

5. A small challenge

 Return to the Describe > One Variable > Graph dialogue.
 With the same 6 variables, from **Depth to z**, change from a boxplot to a Violin Plot (Don't worry that you may not know what a violin plot is).

 Look at the curious shape (Fig. 13) for some of the variables, particularly for the one called **table**. (This is showing something about the data that is not evident from a boxplot.)

Fig. 13 Curious results from a violin plot
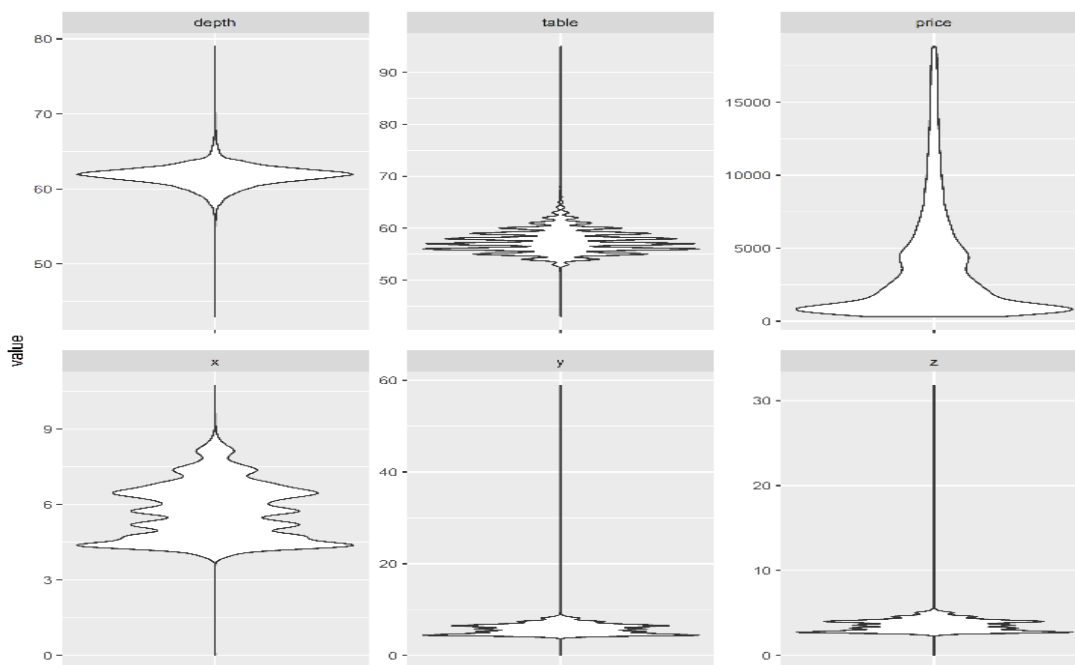
 Examine this further by repeating the violin plot for just the variable called **table**. Now use the dialogue **Describe > One Variable > Frequencies** for the variable **table**. What do you notice?

6. A more ambitious analysis

 Go to the *Describe > Multivariate > Correlations* dialog. (Note that only the numeric columns are visible for this dialog.)
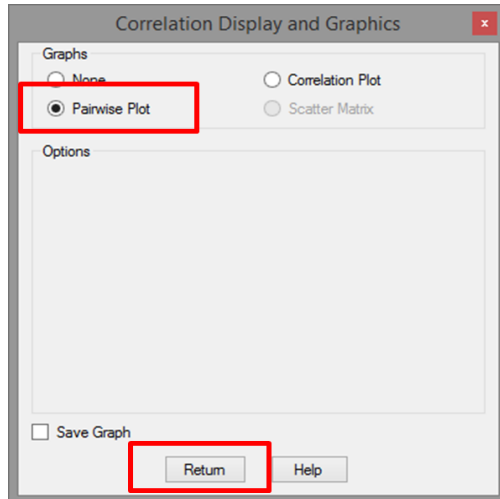 Select the **Multiple Columns** button at the top of the dialogue, Fig. 14.
 Select the first 2 variables (Carat and Depth) and the last two (y and z), Fig. 14.
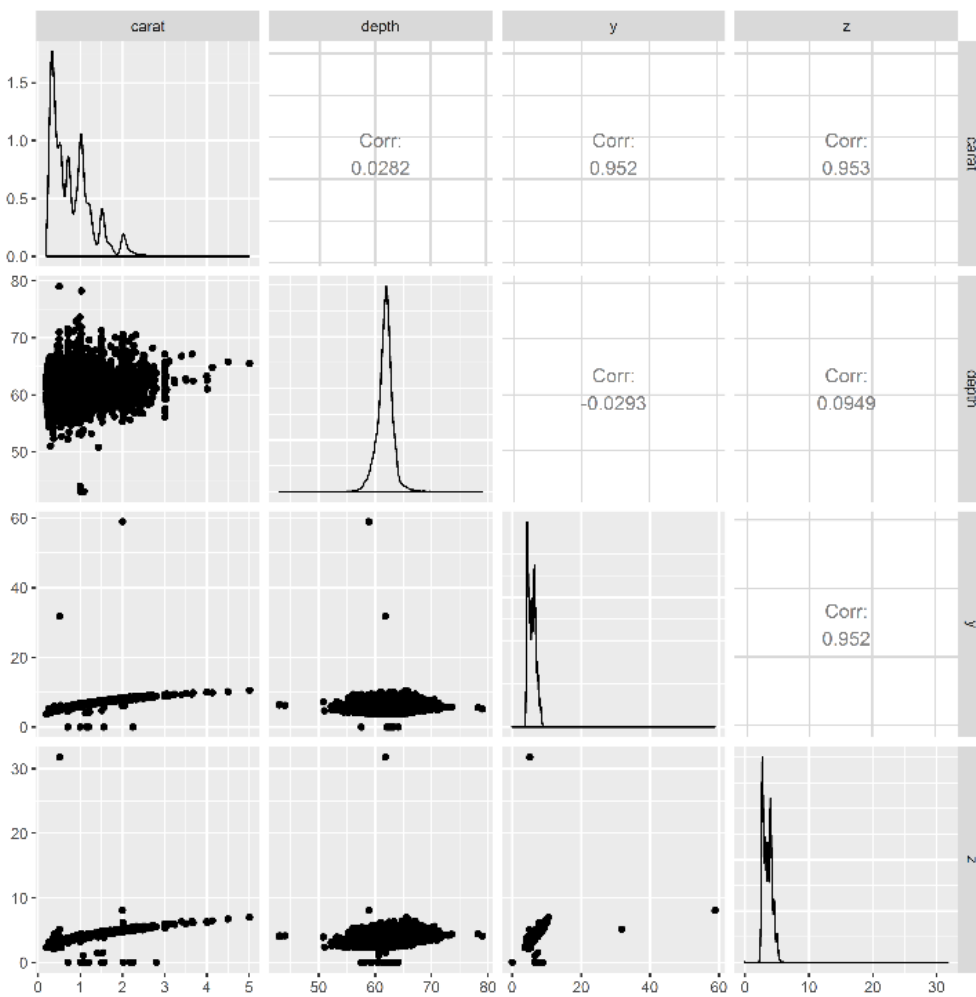 Click on the **Options button** to go to the sub-dialogue, Fig. 14.

Fig. 14. The Correlations dialogue   And sub-dialogue
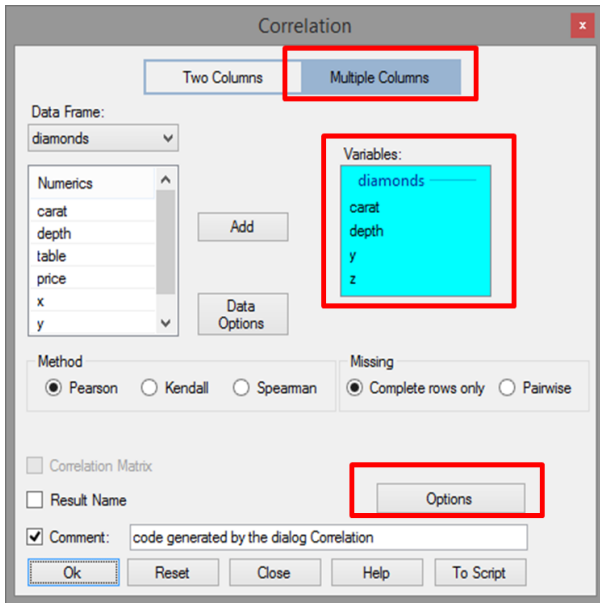
 Select the *Pairwise Plot*. Then press *Return*

 Press *OK* to give the results shown in Fig. 15.

Fig. 15 Correlations

7. Reflections

It is easy to follow instructions without being clear on the main points being covered. We list here some of the points that have been covered:

- **File > Open from Library** was used to choose a data set for analysis. Similarly the **File > Open** dialogue can be used to import your own data.
- The data were well organised and ready for analysis, so we used the **Describe** menu.
- Initial exploration of data often starts by examining variables one at a time. So we started with the **Describe > One Variable > Graph** dialogue.
- In almost every dialog the first step is to **select the variables** for analysis.
- We often had to return to a dialogue to refine the analysis.
- The dialogues "remembered" their last settings, so small changes were quick to do.
- Some dialogues have sub-dialogues that give more options.
- On the statistical side it was very easy to produce "multiple graphs". They are useful.
- Finally we wonder whether you consider Fig. 15 to be a graph or a table?. It has some characteristics of both and the merging of these ideas is one reason we have chosen to distinguish between **Describe** and **Model** in the menus in R-Instat, rather than the more traditional **Graphics** and **Statistics**.

8. Next steps

You can continue exploring the describe menu with this data set and produce more tables and graphs that explore the data. The next part of the tutorial introduces dialogues in the **Prepare** menu using a second data set from the R-Instat library.

9. Feedback and reporting bugs

R-Instat is still under active development with many improvements and new features planned for future versions. We appreciate feedback you can have to help us improve R-Instat. There are several ways you can provide your feedback:

1. For general feedback you can contact us via email at R-Instat@AfricanMathsInitiative.net.
2. Our issues page on our GitHub account can be used to report specific bugs or suggestions and this is the most direct way to contact the development team. Note that our issues page is publicly visible to anyone. It can be accessed here: https://github.com/africanmathsinitiative/R-Instat/issues. Click the green **New Issue** button on the right side to send your message.

When reporting a bug or problem, it's most helpful to us if you can be as specific as possible and detail how to reproduce the bug, pasting the R code from the log file and attaching data if possible.

R-Instat Team, African Data Initiative