INTRODUCTORY TUTORIAL PART 1: DESCRIBING DATA

BY: ROGER STERN, DANNY PARSONS, JAMES MUSYOKA , DAVID STERN AND BERYL JOHNS

March 2021

CONTENTS

CHAPTER 1 — 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.

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.

You can find out more about the ADI (R-Instat) Team at:

R-Instat@AfricanMathsInitiative.net

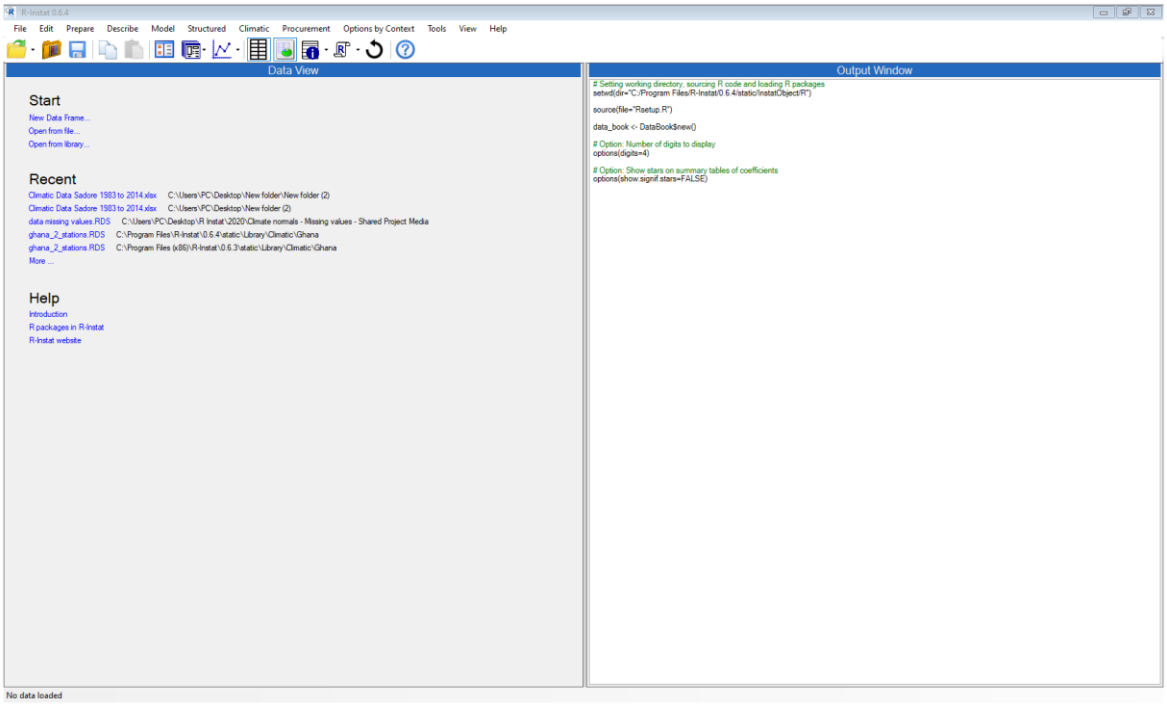
CHAPTER 2 — EXPLORING R-INSTAT

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

2.1 THE INSTALLATION

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

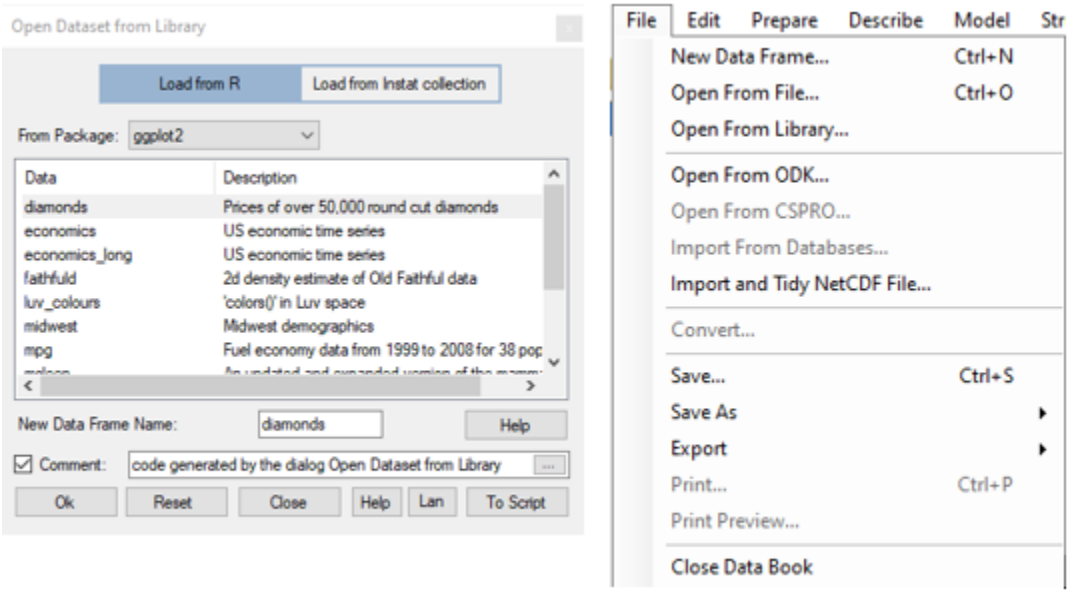
Fig 1: R-Instat main Interface



2.2 A FIRST TASK – IMPORTING DATA FROM THE LIBRARY

- Go to ***File > Open From Library*** or click on the ***Open from library…*** shortcut
- Click on the ***From Package*** dropdown and choose ***ggplot2***.
- Choose the first example, ***diamonds*** as shown in Fig. 2.

Fig. 2. Using a library dataset



- 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, though you do not need an internet connection to view it. (It is the dataset used by Hadley Wickham, the author of ggplot2, for many of the examples in his own documentation.)
- Now return to the dialog, select the ***diamonds*** dataset again and press ***OK***.

Fig. 3 The diamonds data

Data View

	carat	cut (o.f)	color (o.f)	clarity (o.f)	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Goo	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Goo	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Goo	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Goo	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42	2.68
17	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34	2.68
18	0.30	Good	J	SI1	63.4	54.0	351	4.23	4.29	2.70
19	0.30	Good	J	SI1	63.8	56.0	351	4.23	4.26	2.71

diamonds

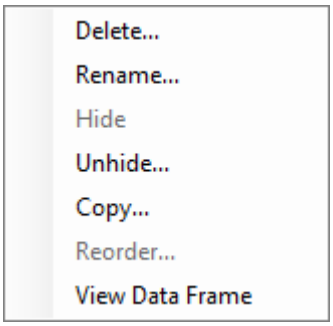
Showing 1000 of 53940 rows | Showing 10 of 10 columns

diamonds

The data appear that on the left-hand side – looks a little like a spreadsheet.

- Scroll to the bottom of the data to see it appears to have just 1000 rows. This is just a small window onto the full data set of over 53 thousand rows.
- To see the full data set, ***right click on the bottom tab***, Fig. 4.
- Choose the last option ***View Data Frame***, also shown in Fig. 4.

Fig. 4. Viewing a data set



Data: diamonds

File

	carat	cut	color	clarity	depth	table	price	x	y
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75
16	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42
17	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34
18	0.30	Good	J	SI1	63.4	54.0	351	4.23	4.29
19	0.30	Good	J	SI1	63.8	56.0	351	4.23	4.26

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, Fig 5 .

Fig. 5. Variables

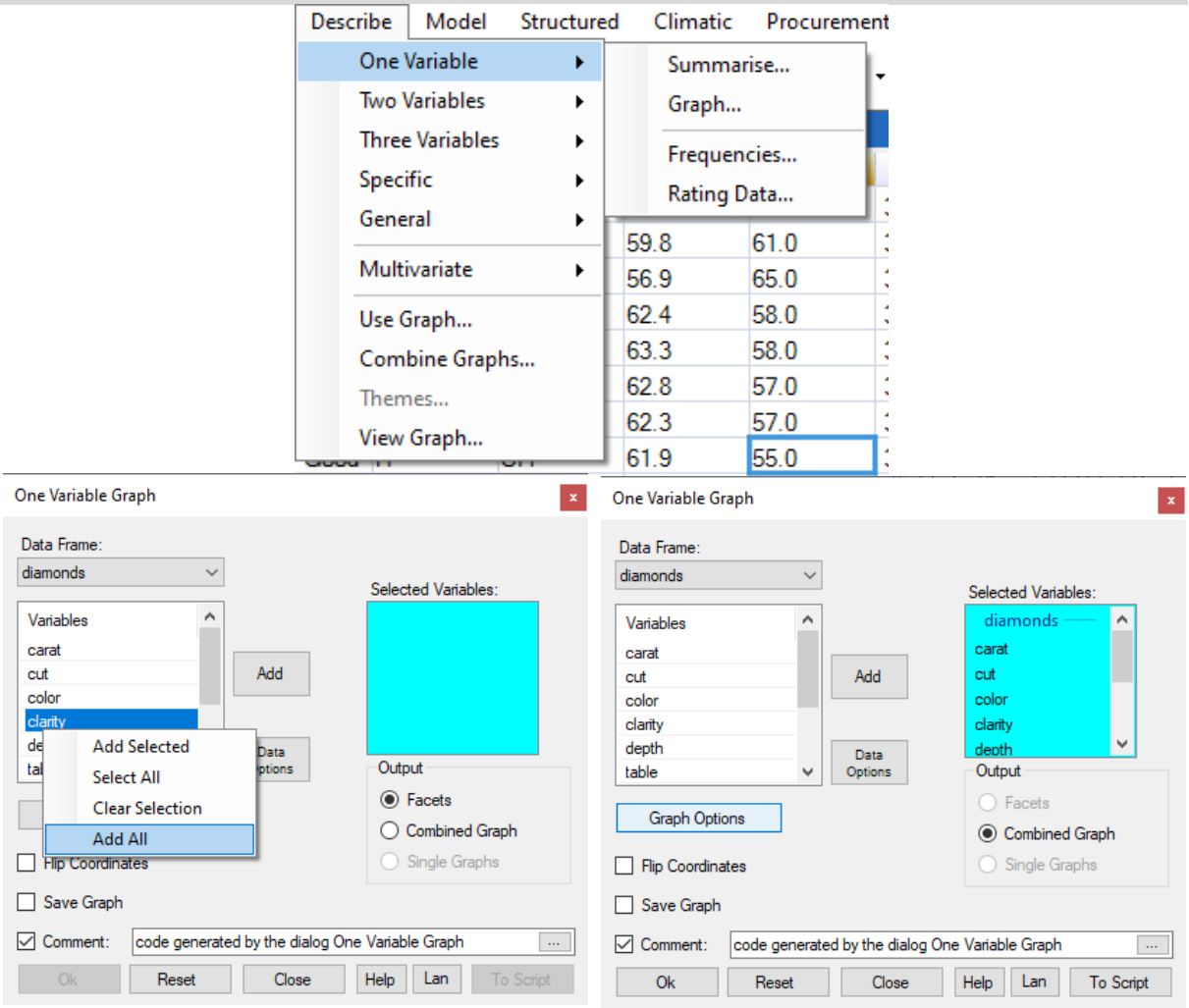
	carat	cut (o.f)	color (o.f)	clarity (o.f)	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63

Next we would usually spend some time organizing the data using the **pepare menu**, here our data is already well prepared so we will go straight to the analysis starting with the **describe menu**.

2.3 SOME GRAPHS

- Go to ***Describe > One Variable > Graph***, Fig. 6.
- ***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. 6.

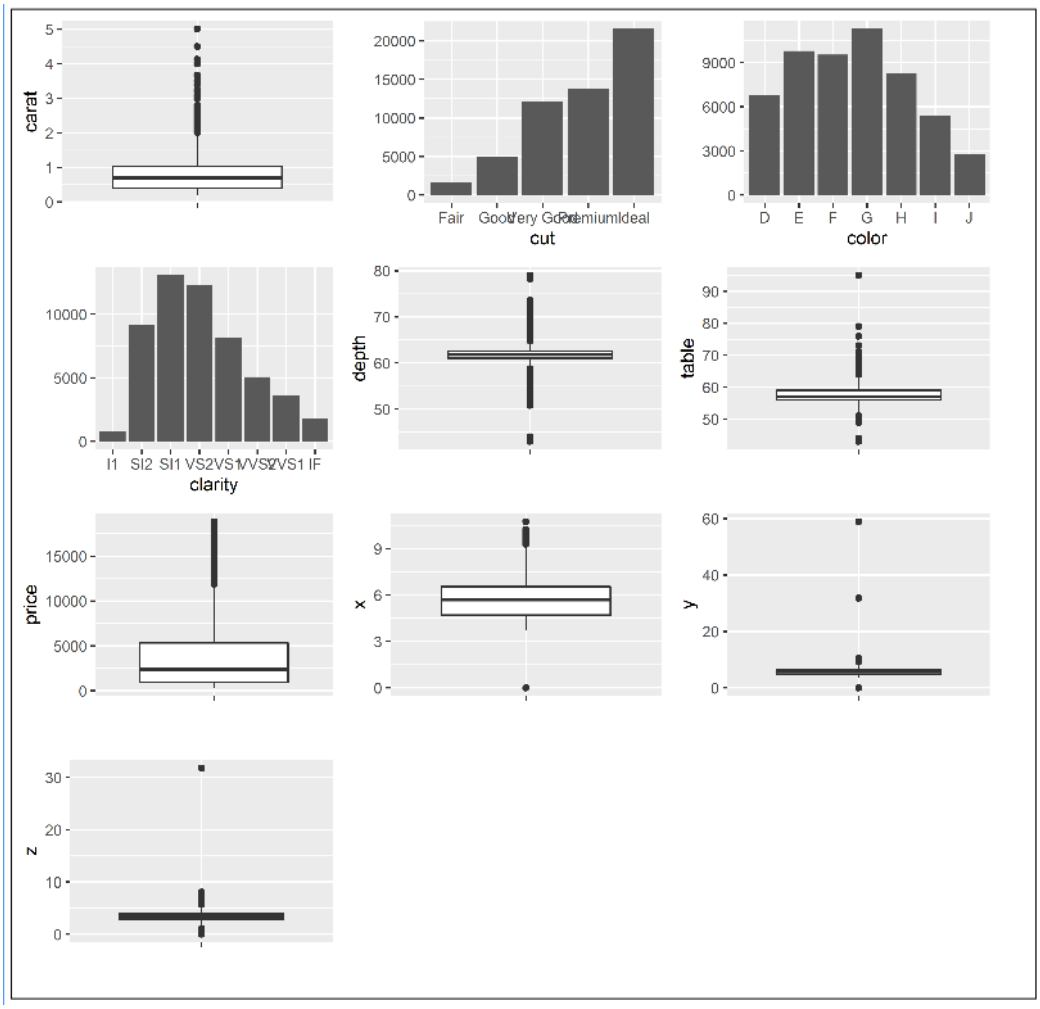
Fig 6: One variable graphs dialog



In the dialog in Fig. 6 the radio button changed from ***Facets*** to ***Combine Graph***, see 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. 7.

Fig 7: One Variable Graphs

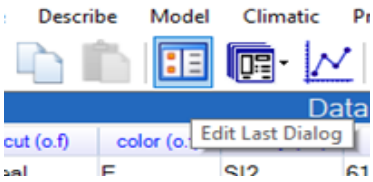


The analysis in R has detected which variables are categorical and given bar charts for them, compared with boxplots for the numeric variables.

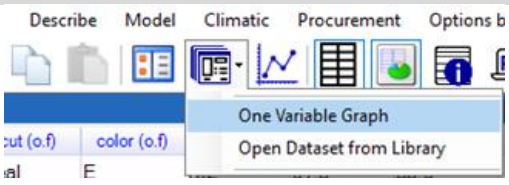
Often, the results from using a dialogue can be improved, so you wish to use it again. You could use the same menu options as in Fig. 6, but there is a quicker way.

- Click on the little **dialogue icon** on the toolbar, see Fig. 8, which takes you back to the **previous dialogue**. (Or the next icon that lets you return to any of the **last 10** recently used dialogues.)

Fig 8: Use the toolbar to return to a dialog



Or any of the recent dialogs



You see the dialogue has "remembered" the settings just as you left it when you pressed OK. This is often convenient.

- But this time press the **Reset** button at the bottom of the dialogue, to clear all the settings.
- Then select the **last 6 variables**, to put into the receiver.

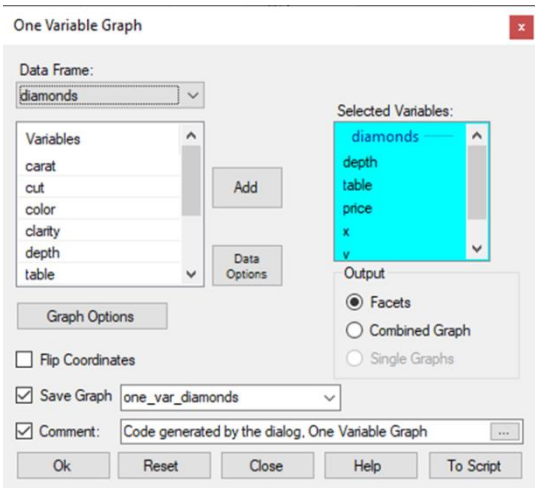
As these are all numeric columns the radio buttons on the right stays on Facets, so you can see what this is!

- 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**

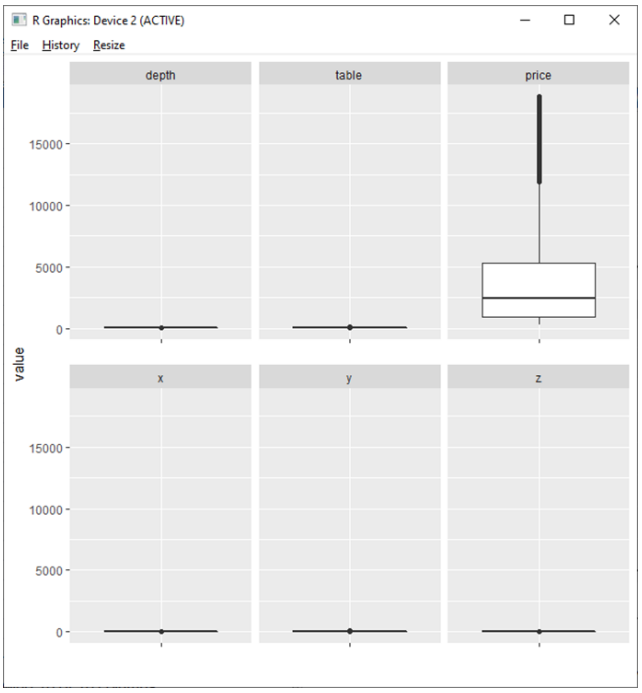
The dialogue didn't work. Instead, it gives a message that "The name cannot contain a space" (or a dash). It is the name of an object in R and these are not allowed.

- Click on **OK** to clear the message box.
- Change the name to **OneVarDiamonds** or perhaps **one_var_diamonds**, Fig. 9, and click **OK** again.

Fig 9: The One Variable Graph dialog again



With a faceted graph



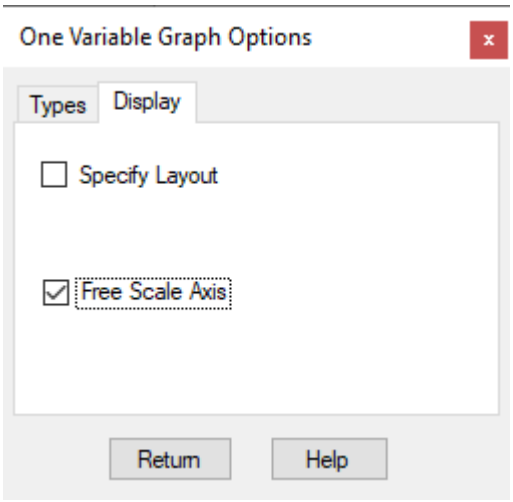
This shows a **faceted** graph, Fig. 9. It doesn't look very nice. Can you see why? By default, the y-axis is the same for all the graphs. This is often what is wanted for a multiple graph because then you don't then need the axis to be labelled for each variable. However, it isn't what we need here. The different variables have very different scales, and we need to reflect this in the graph. There is an easy fix.

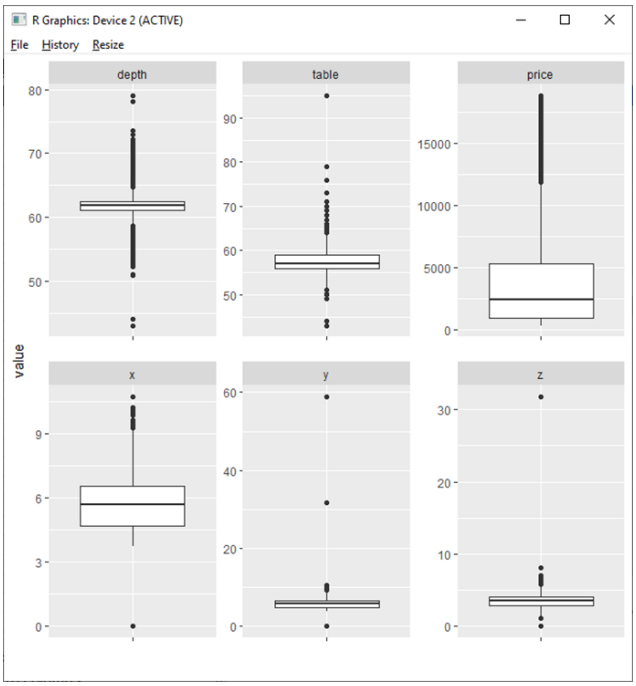
- Return to the same dialogue again.
- Click on the **Graph Option**s button.

You now see a sub-dialogue with just 2 tabs, Fig. 10. One tab allows you to change the type of graph that is shown, so you could use this to get a histogram instead of a boxplot for example. The other changes how the graph is displayed.

- Press on the tab labelled **Display** and then click on the **Free Scale Axis**.
- Press on the **Return** button and then **OK** again, to give the graph also shown in Fig. 10.

Fig 10: The One Variable graph sub dialog The next graph





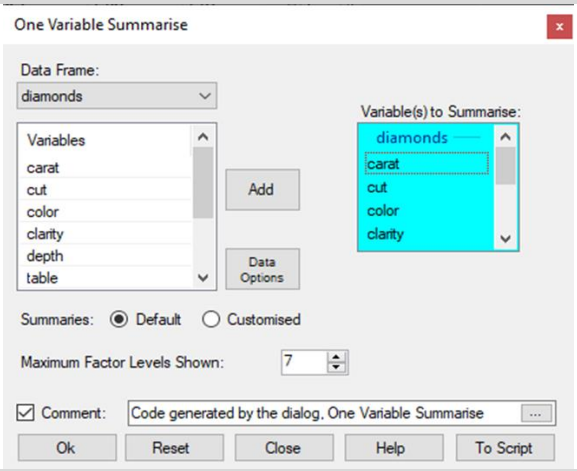
Now our boxplots are clear, and each y-axis has its own scale.

2.4 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. 11.
- Press **OK** to give the results also shown in Fig. 11.

Fig 11: The One Variable Summarise dialog



With some results

```
     carat               cut        color        clarity          depth      
 Min.   :0.200   Fair      : 1610   D: 6775   SI1     :13065   Min.   :43.0  
 1st Qu.:0.400   Good      : 4906   E: 9797   VS2     :12258   1st Qu.:61.0  
 Median :0.700   Very Good:12082   F: 9542   SI2     : 9194   Median :61.8  
 Mean   :0.798   Premium   :13791   G:11292   VS1     : 8171   Mean   :61.8  
 3rd Qu.:1.040   Ideal     :21551   H: 8304   VVS2    : 5066   3rd Qu.:62.5  
 Max.   :5.010                     I: 5422   VVS1    : 3655   Max.   :79.0  
                                   J: 2808   (Other): 2531                  
     table           price             x                y                z          
 Min.   :43.0   Min.   :  326   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:56.0   1st Qu.:  950   1st Qu.: 4.71   1st Qu.: 4.72   1st Qu.: 2.91  
 Median :57.0   Median : 2401   Median : 5.70   Median : 5.71   Median : 3.53  
 Mean   :57.5   Mean   : 3933   Mean   : 5.73   Mean   : 5.73   Mean   : 3.54  
 3rd Qu.:59.0   3rd Qu.: 5324   3rd Qu.: 6.54   3rd Qu.: 6.54   3rd Qu.: 4.04  
 Max.   :95.0   Max.   :18823   Max.   :10.74   Max.   :58.90   Max.   :31.80  
```

This is almost right, but the variable 'clarity', marked with 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 into a category called (Other). This can also be easily corrected.

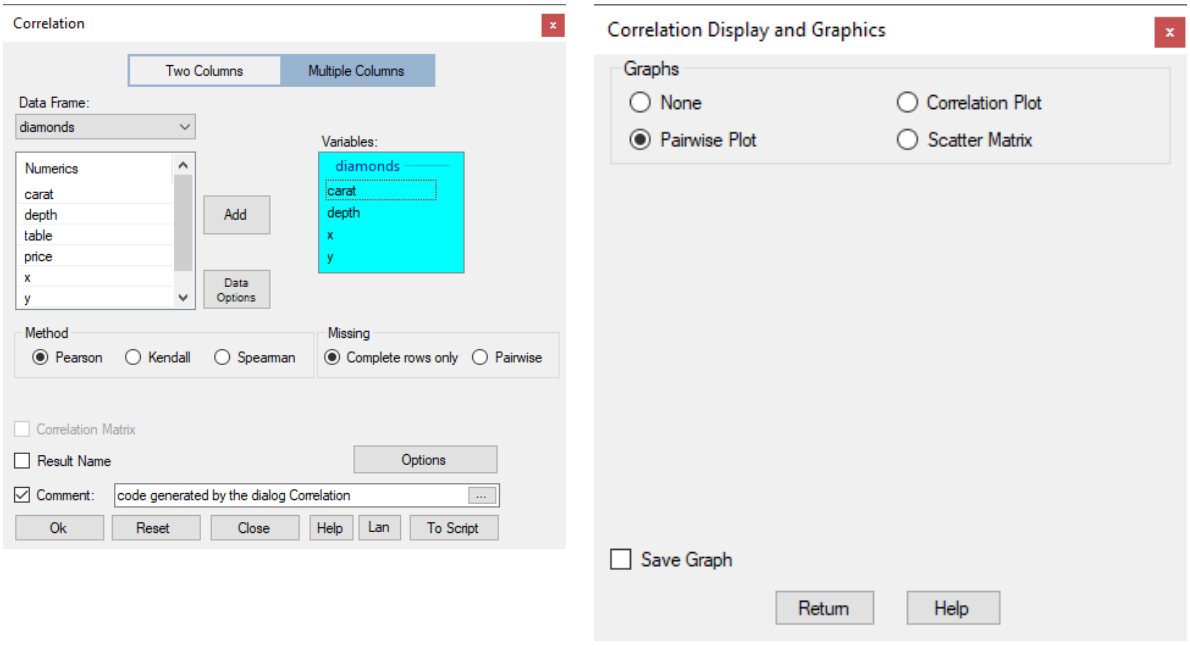
- Return to the last dialogue.
- In the dialogue, Fig. 11, change the **Maximum Factor Levels Shown** from 7 to 10. Press **OK**.

The levels are now all given for the **clarity** factor column.

2.5 A MORE AMBITIOUS ANALYSIS

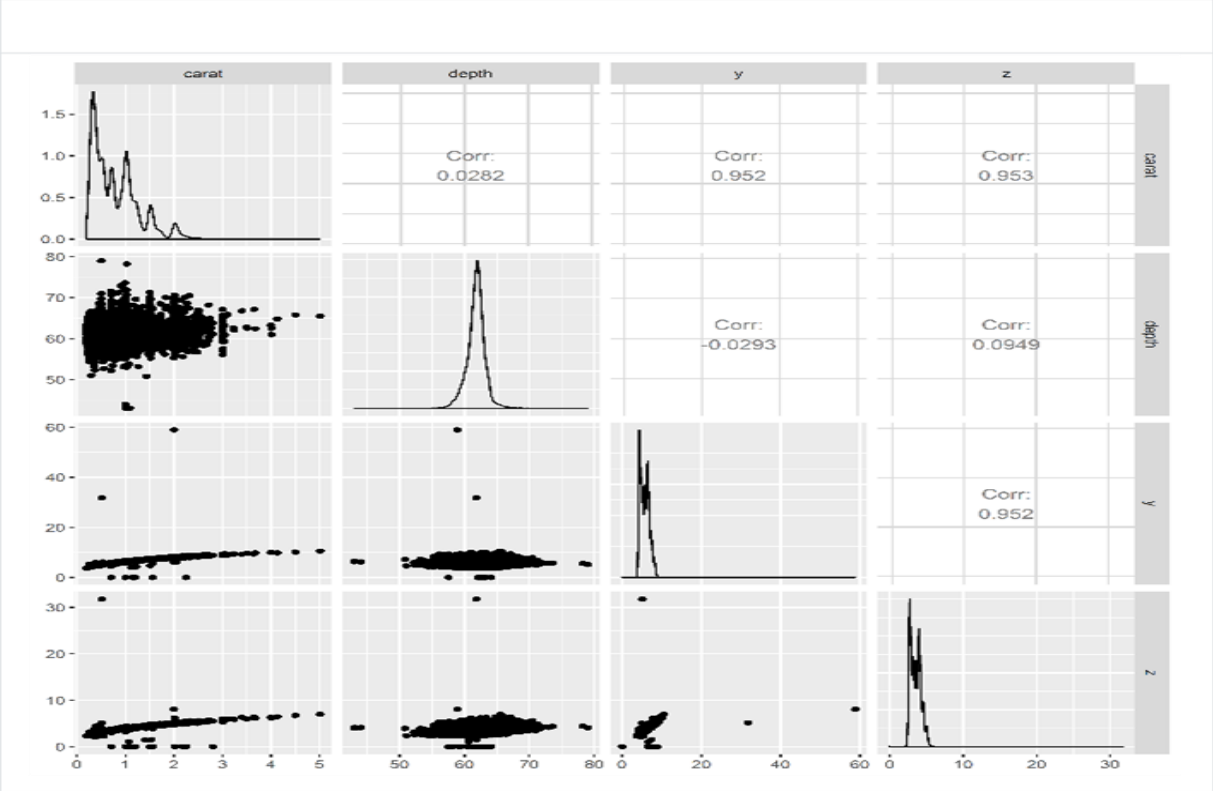
- Go to the ***Describe > Multivariate > Correlations*** dialog. (Note that only the 7 the numeric columns are visible in this dialog.)
- Select the **Multiple Columns** button at the top of the dialogue, Fig. 12.
- Select the first 2 variables (**Carat and Depth**) and the last two (**y and z**), Fig. 12.
- Click on the **Options button** to go to the sub-dialogue, Fig. 12.
- Select the ***Pairwise Plot***. Then press ***Return***

Fig 12: The correlations dialog And sub dialog



- Press ***OK*** to give the results shown in Fig. 13.

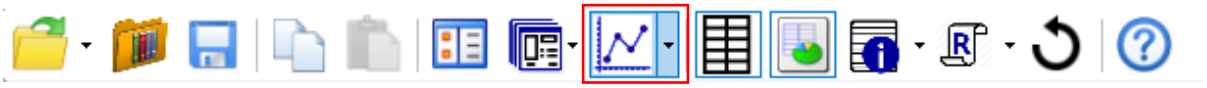
Fig 13: Correlations



Once you have the graph, it might be clearer in the R viewer, so click on the graph icon in the toolbar.

➢ Click on the graph icon in the toolbar Fig. 14.

Fig 14: Correlations



CHAPTER 3 — 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.

CHAPTER 4 — 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.

CHAPTER 5 — 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