R-INSTAT INTRODUCTORY TUTORIAL 1

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

November 2022

CONTENTS

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 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 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:

http://r-instat.org/

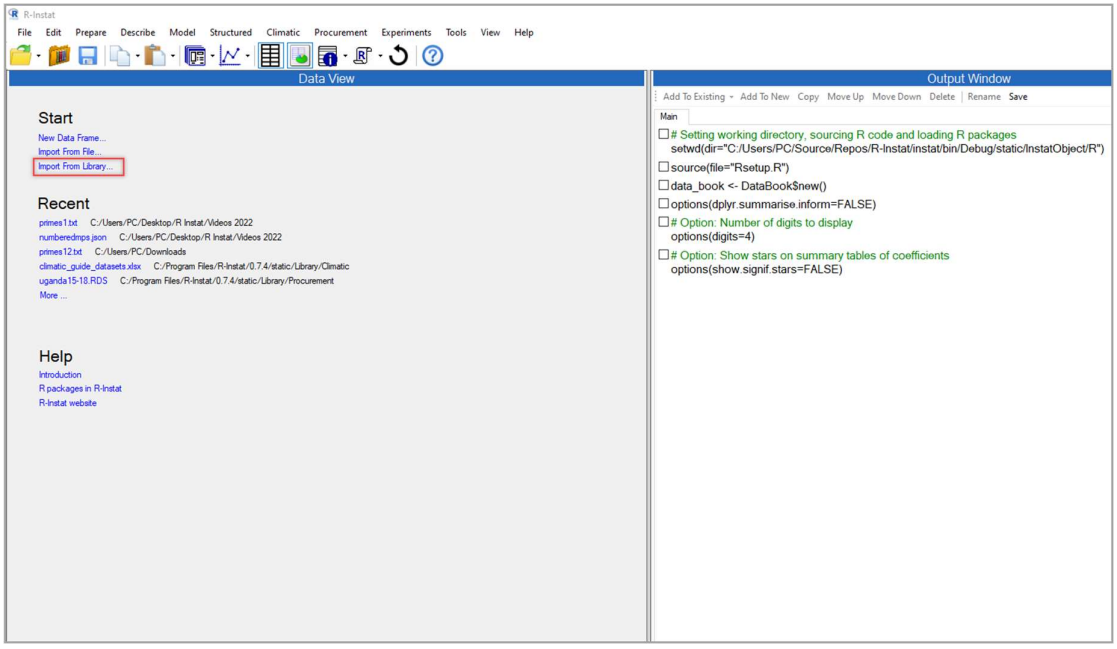
2. EXPLORING R-INSTAT

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

STARTING

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

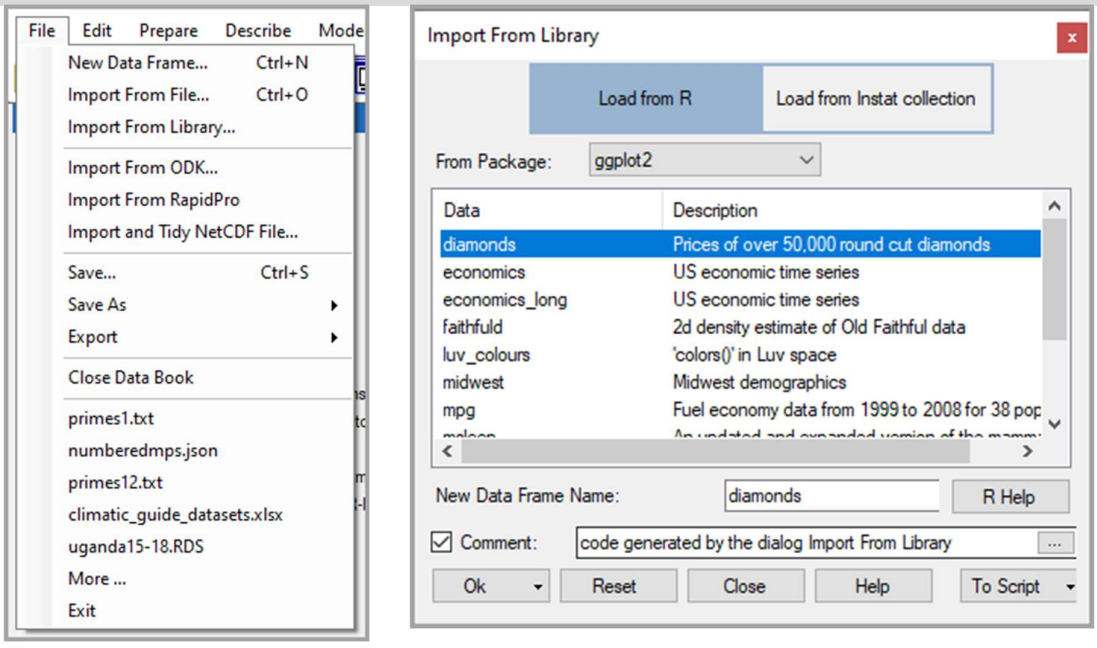
Fig. 1: R-Instat main Interface



A FIRST TASK – IMPORTING DATA FROM THE LIBRARY

- Go to ***File > Open From Library*** or click on the ***Open from library…*** shortcut in Fig. 1.
- 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



- A second ***Help*** button is now enabled, just below the list of datasets. Click on that button to get further information about the dataset. (The diamonds dataset is used by Hadley Wickham, the author of ggplot2, for many examples in his own documentation.)
- Now return to the dialog, select the ***diamonds*** dataset again and press ***OK***.

Fig. 3 The diamonds data

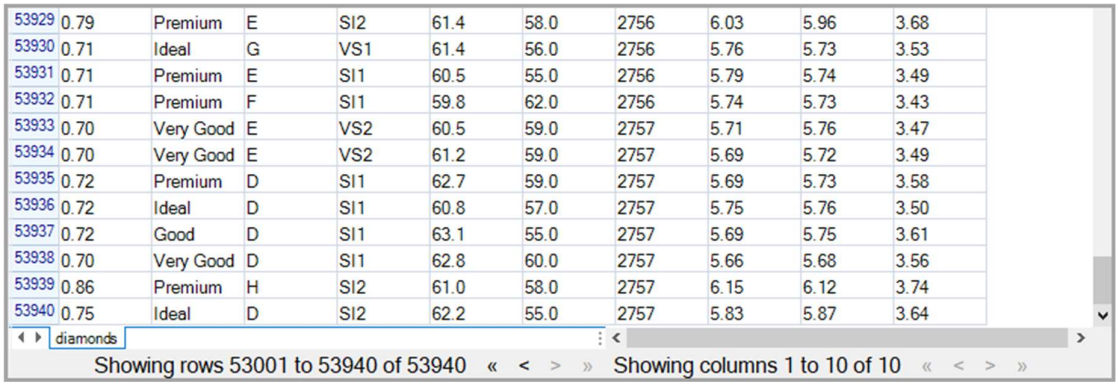
Data View

	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 Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	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 Good	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

The data, in Fig. 3, appear that on the left-hand side – looks a little like a spreadsheet.

- Scroll to the bottom of the data to see that it is currently showing the first 1000 of 53,940 rows.
- Use the single arrow to move to the next thousand rows.
- Or the double arrow to go straight to the end of the data set, 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 ordered, for example the second column, namely the **cut** of the diamond's 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

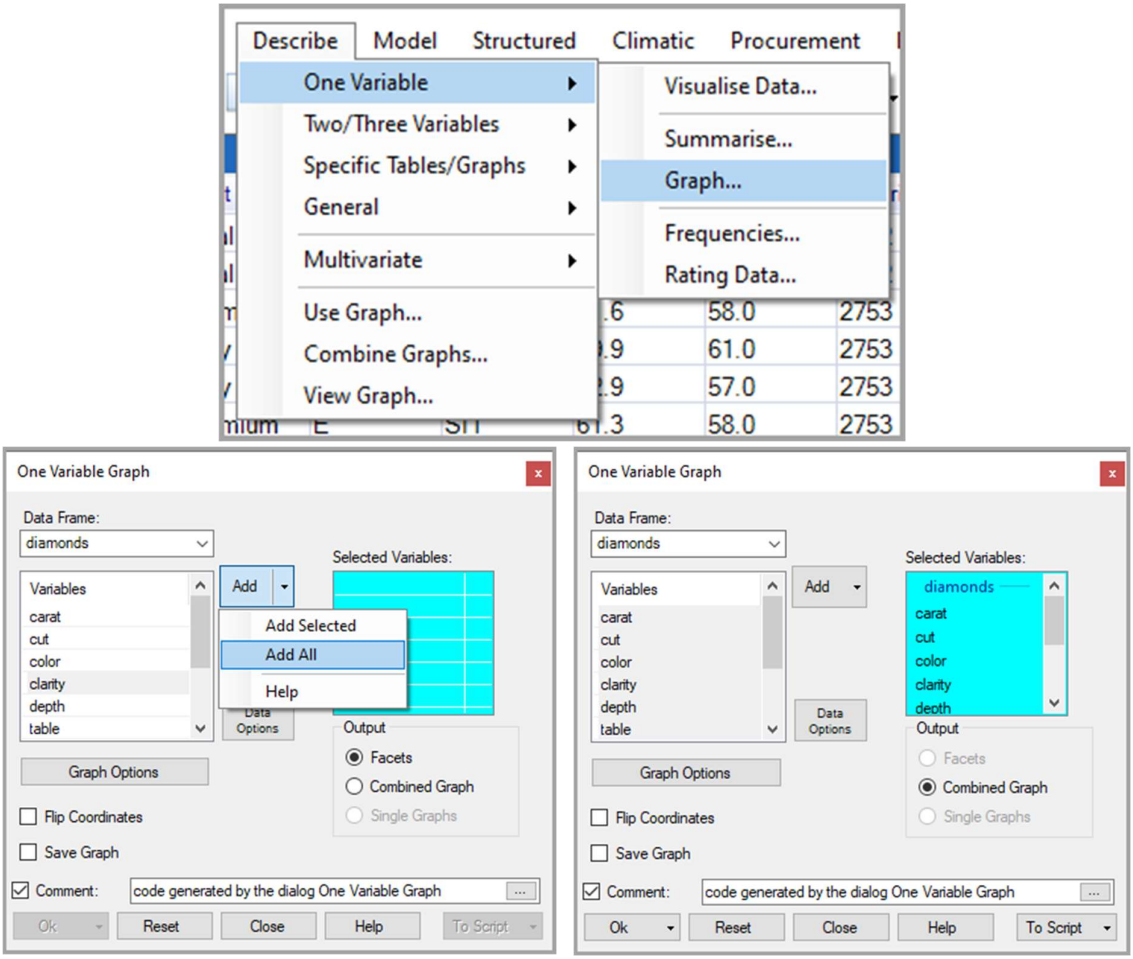
	carat	cut (O.F)	color (O.F)	clarity (O.F)	depth	table	price	x	y	z
53900	0.72	Ideal	H	VVS2	62.3	56.0	2752	5.74	5.81	3.60
53901	0.73	Ideal	H	VS2	62.5	58.0	2752	5.71	5.75	3.58
53902	0.57	Premium	E	VS1	61.6	58.0	2753	5.36	5.33	3.29

Next, we usually use the ***prepare menu*** to organize the data. ***But here*** our data is already well prepared so we go straight to the ***describe menu*** to start the analysis.

SOME GRAPHS

- Go to ***Describe > One Variable > Graph***, Fig. 6.
- ***Click*** on the down arrow of the Add button and select **Add All**. (Or just select all the columns and then click ***Add***)

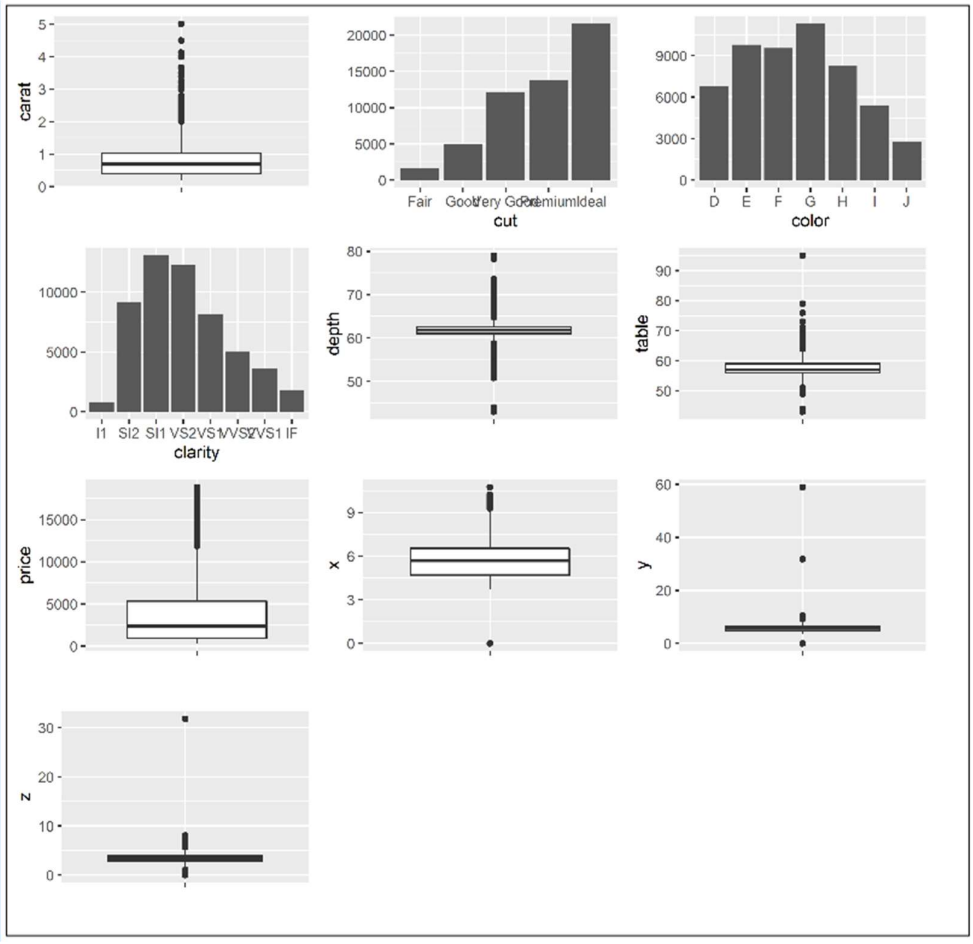
Fig 6: One variable graphs dialog



In the dialog in Fig. 6 the radio button changed from ***Facets*** to ***Combine Graph*** That is because the selected variables are of different data types; some are categorical (factors) while others are numeric.

- Press ***OK*** to give the results also shown in Fig. 7.

Fig. 7: One Variable Graphs

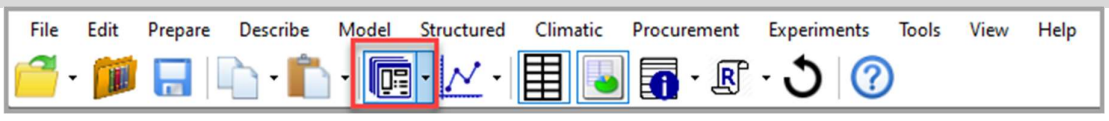


The analysis in R has detected which variables are categorical and shown them as bar charts, compared with boxplots for the numeric variables.

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

- Use the **dialogue icon** on the toolbar, see Fig. 8. A double click takes you back to the **previous dialogue**. Or the down arrow brings up a dropdown list of the **last 10** recently used dialogues.

Fig. 8: Use the toolbar to return to a dialog



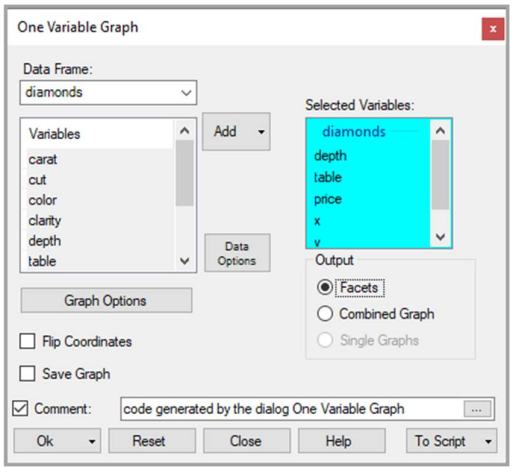
You see the dialogue has "remembered" the settings just as you left it when you pressed OK. This is often convenient.

- This time press the **Reset** button at the bottom of the dialogue, to clear all the settings.
- Then select the **last 6 variables**, so starting with depth, to put into the receiver.

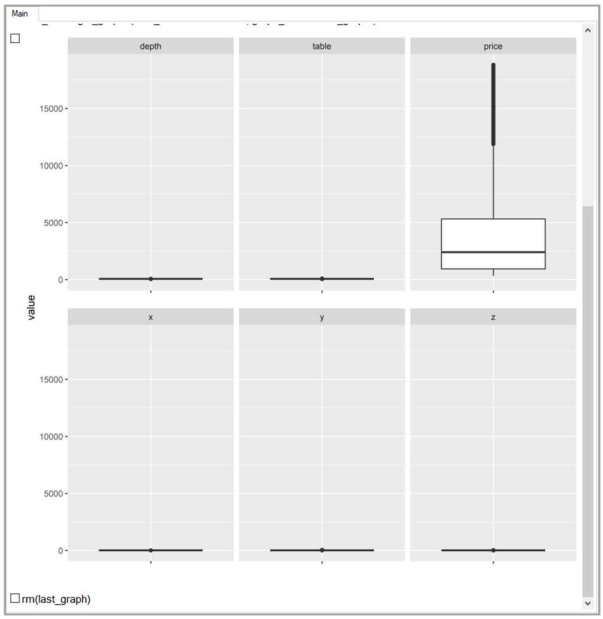
As these are all numeric columns the radio buttons on the right stays on facets, so you can see what they are!

- Now click **OK**

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 you don't then need the axis to be labelled for each variable.

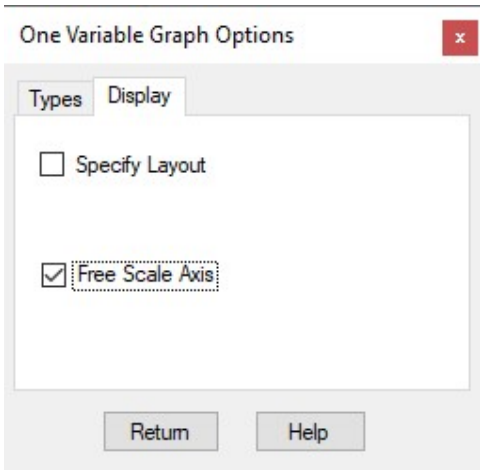
However, it isn't what we need here. The different variables have very different scales, and we need to reflect this in the graph. There is an easy fix.

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

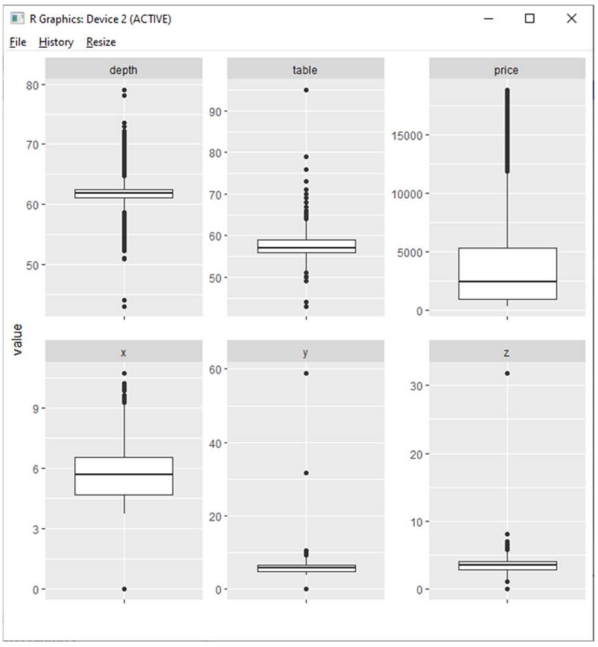
You now see a sub-dialogue with just 2 tabs, Fig. 10. One tab allows you to change the type of graph that is shown, so you could use this to get a histogram instead of a boxplot for example. The other tab changes how the graph is displayed.

- 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



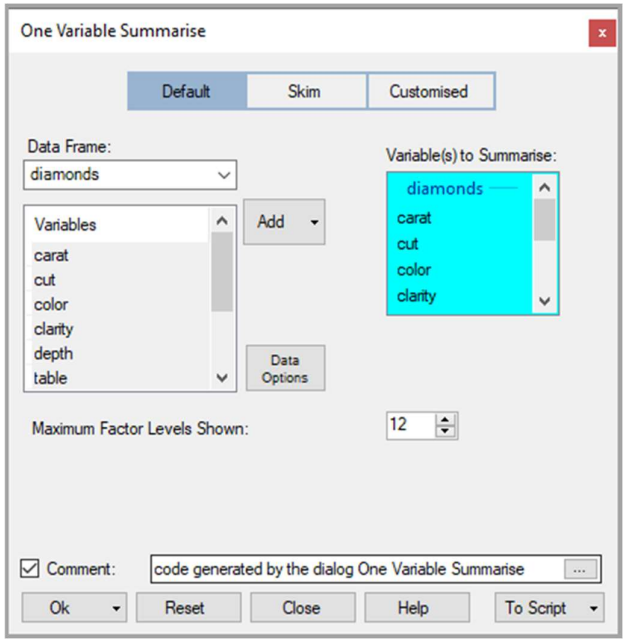
Now our boxplots are clear, and each y-axis has its own scale.

SOME SUMMARIES

Often analyses involve numerical as well as graphical summaries.

- Go to ***Describe > One Variable > Summarise***.
- This dialog has three tabs, that summarise the data in different ways. We use the Default view, Fig. 11.
- Select all the variables again (as you did with for the first use of the Graph dialogue).
- Press **OK** to give the results also shown in Fig. 11.

Fig 11: The One Variable Summarise dialog



With some results

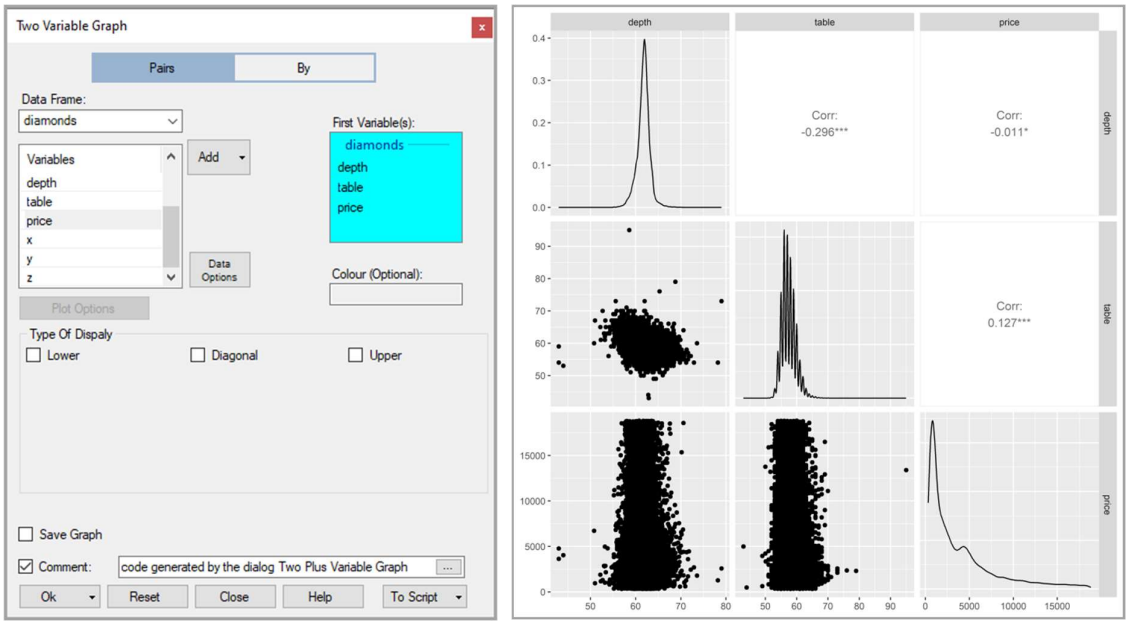
```
     carat               cut        color     clarity          depth
 Min.   :0.200   Fair      : 1610   D: 6775   I1   :  741   Min.   :43.0
 1st Qu.:0.400   Good      : 4906   E: 9797   SI2  : 9194   1st Qu.:61.0
 Median :0.700   Very Good:12082   F: 9542   SI1  :13065   Median :61.8
 Mean   :0.798   Premium   :13791   G:11292   VS2  :12258   Mean   :61.8
 3rd Qu.:1.040   Ideal     :21551   H: 8304   VS1  : 8171   3rd Qu.:62.5
 Max.   :5.010                     I: 5422   VVS2 : 5066   Max.   :79.0
                                   J: 2808   VVS1 : 3655
                                             IF   : 1790
     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
```

These results, just like the graph, are different for the numeric and factor variables. For the numeric variables they provide the mean together with the same 5 summaries (Min, 1st Qu., Median, 3rd Qu. And Max) used in a boxplot. For a factor, the frequencies at each level are given, as we saw earlier in the bar charts.

A MORE AMBITIOUS ANALYSIS

- Go to the ***Describe > Two Variables > Graph*** dialog.
- First use the 3 variables: depth, table and price. Then click OK to show the graph.

Fig 12: Two Variable Graph



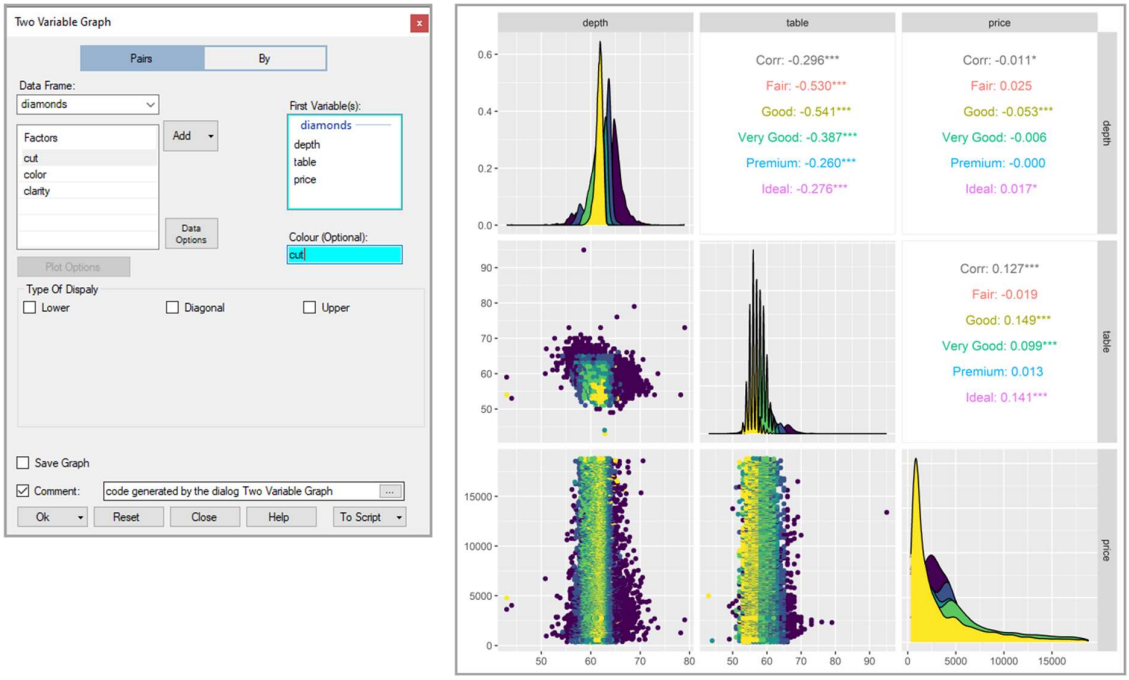
The results show the density of each variable on the diagonal. The 3 correlations are on the upper part of the display, while the 2-variable scatterplots are on the lower part.

- Return to the dialogue to create a second graph. You do this by double clicking on the ***View/Edit Last Dialog*** icon.

Fig 13: View/Edit Last Dialog Icon

- Add the ***Cut*** factor to the ***Colour (Optional)*** receiver.

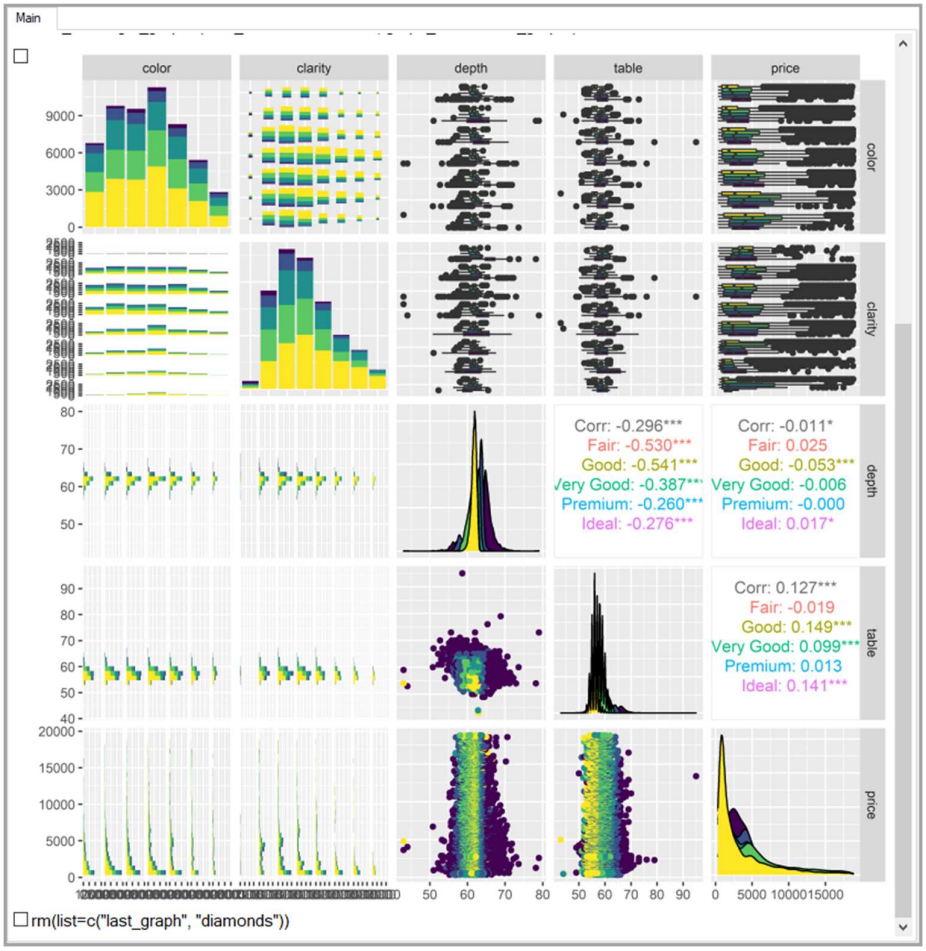
Fig 14: Two variable Graph 2



It looks quite nice in colour! And now the correlations, and the scatterplots are given separately for each level of the cut factor.

- Let's make one more graph. Return to the dialog again, either through the menus or using the ***View/Edit Last Dialog*** icon.
- This dialog allows both factors and numeric variables. We add the factor variables colour and clarity. What do we get then?

Fig 15: Colour Graph with Numeric and Factor Variables



Once you have the graph, it might be clearer maximised

- Click on the ***blue maximise*** link in the ***Output window*** Fig. 16

Fig. 16: Maximise

CORRELATIONS

The correlations were given as a sort of extra in the ***Two-Variables > Graph*** dialogue above. We now look at more correlations and in more detail.

We examine the relationship between the data in the x and z variables. The diamond data R-Help file, accessed earlier via the Import From Library dialog, explained that:

- x - length in mm (0–10.74)
- z - depth in mm (0–31.8)

➢ Go to the **Describe > Two/Three Variables > Correlations** (Fig. 17)

➢ Put the **variables x** and **z** into the 2 receivers. R-Instat needs to know both variables before doing a correlation, so the **OK** button is not be enabled until they are entered.

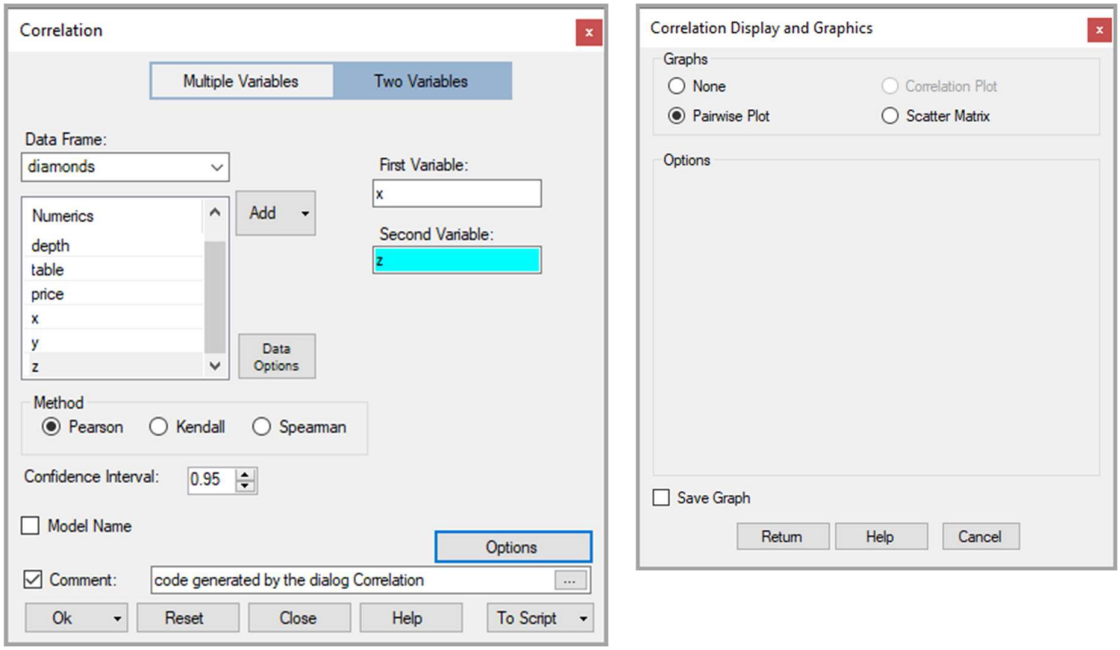
This dialog has an ***Options*** sub-dialogue (Fig. 17) to add further options choices.

➢ Click on it. Select the **Pairwise Plot**, then click return.

Back on the main dialog.

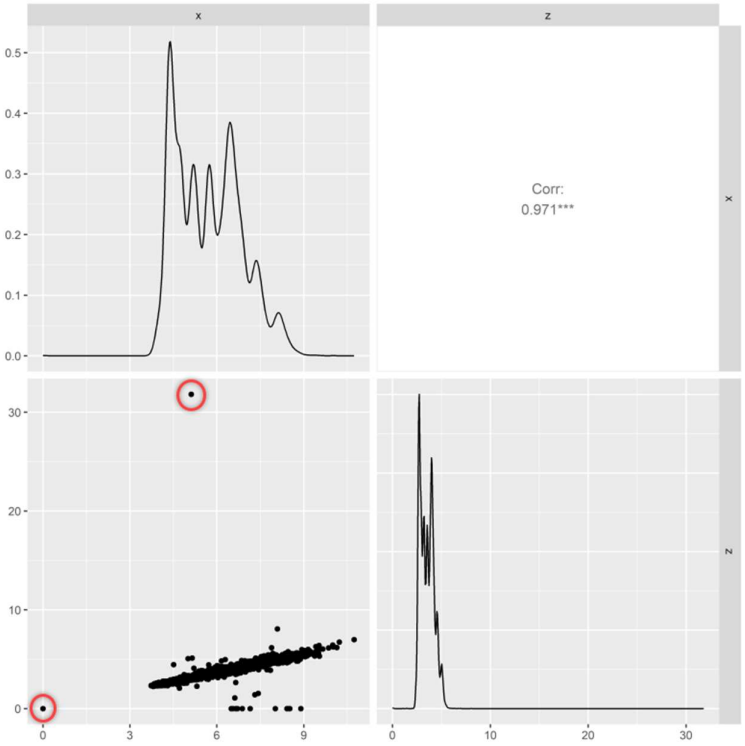
➢ Click **Ok**

Fig. 17: Correlations Dialog and Sub-Dialog



The resulting pairwise plot (Fig. 18) shows the correlation between the length and depth of the diamonds is very high, which is not surprising.

Fig. 18: Pairwise Plot



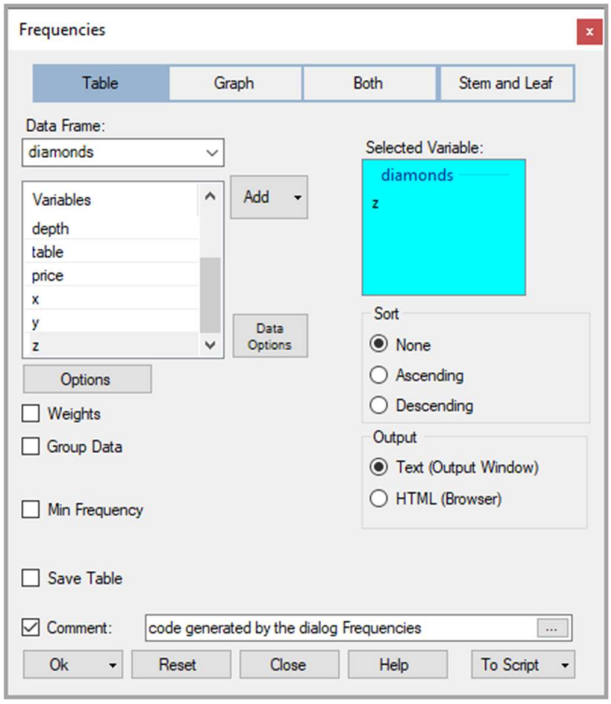
But there are some very odd values, these have been circled in red in Fig. 18. It looks as though some diamonds have a depth of zero, which is obviously impossible and there is at least one value that is 3 times larger than any other. With 53 thousand values it is perhaps not surprising that there are some oddities to investigate.

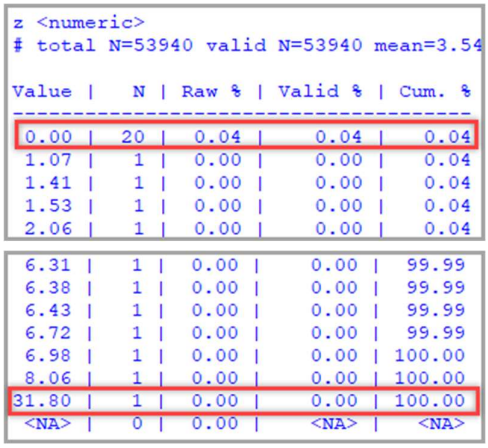
OUTLIERS

For this investigation,

- Go to the ***Describe > One Variable > Frequencies*** (Fig. 19)
- Select the ***z*** variable.
- ***Ok*** is now available, so press it.

Fig. 19: Frequencies Dialog and Results





z <numeric>
total N=53940 valid N=53940 mean=3.54

Value	N	Raw %	Valid %	Cum. %
0.00	20	0.04	0.04	0.04
1.07	1	0.00	0.00	0.04
1.41	1	0.00	0.00	0.04
1.53	1	0.00	0.00	0.04
2.06	1	0.00	0.00	0.04
6.31	1	0.00	0.00	99.99
6.38	1	0.00	0.00	99.99
6.43	1	0.00	0.00	99.99
6.72	1	0.00	0.00	99.99
6.98	1	0.00	0.00	100.00
8.06	1	0.00	0.00	100.00
31.80	1	0.00	0.00	100.00
<NA>	0	0.00	<NA>	<NA>

Examining the results (Fig. 19), shows 20 diamonds have a z, or depth, of zero, and no others are less than 1mm. This is difficult to understand. At the other end the second largest diamond is 8mm deep, then one is 31.80. This is very odd. Perhaps it should have been 3.18?

What to do about outliers is an important topic in data analysis. Here we look briefly at the results without them, to see if excluding them is a good solution.

The ability to filter data is an important and powerful facility in R and hence in R-Instat. It is a slightly more complicated dialogue.

- ***Right-click*** and choose ***Filter***.
- In the ***Filter*** dialog (Fig. 20) choose to ***Define New Filter*** (Fig. 21).
- In the sub-dialogue choose ***z*** and then give the condition as ***> 0*** and ***add the condition***.
- Then use ***z*** again, put ***< 10*** and ***add this condition*** as well.
- Then ***return*** and accept as a filter by pressing ***Ok***.

Fig. 20: Filter Dialog

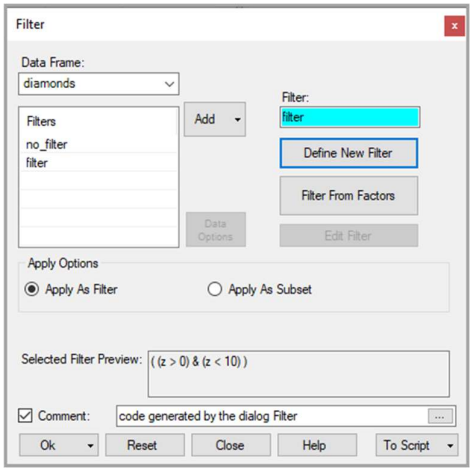
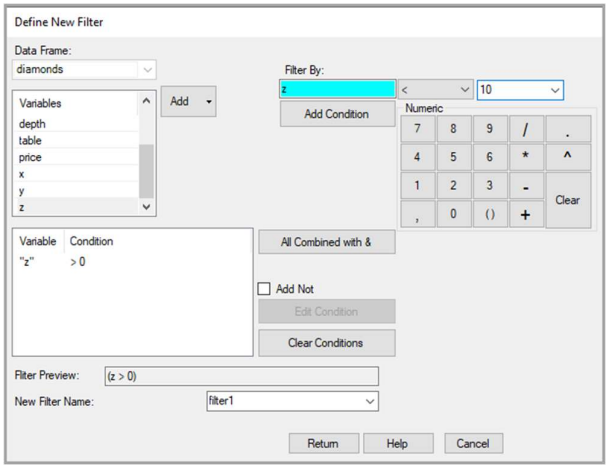
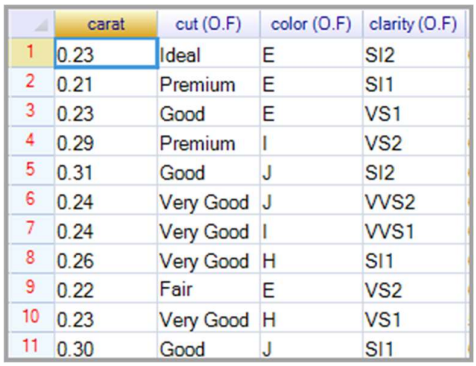


Fig. 21: Sub Dialog

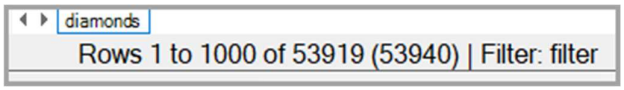


The row numbers on the left, in Fig. 22, are now in red to indicate a filter is in operation. At the bottom the numbers indicate there are now 21 values less – as expected from the results earlier.

Fig. 22: Filter Applied



	carat	cut (O.F)	color (O.F)	clarity (O.F)
1	0.23	Ideal	E	SI2
2	0.21	Premium	E	SI1
3	0.23	Good	E	VS1
4	0.29	Premium	I	VS2
5	0.31	Good	J	SI2
6	0.24	Very Good	J	VVS2
7	0.24	Very Good	I	VVS1
8	0.26	Very Good	H	SI1
9	0.22	Fair	E	VS2
10	0.23	Very Good	H	VS1
11	0.30	Good	J	SI1



- Return to the ***correlation dialogue***. It remains as before, which is convenient, so simply press ***Ok***.

The results look better, though there remain several more possible outliers to investigate. The change in the display from omitting so few observations, shows the importance of paying attention to checking the data quality, before doing too much in an analysis.

Fig. 23: Pairwise Plot

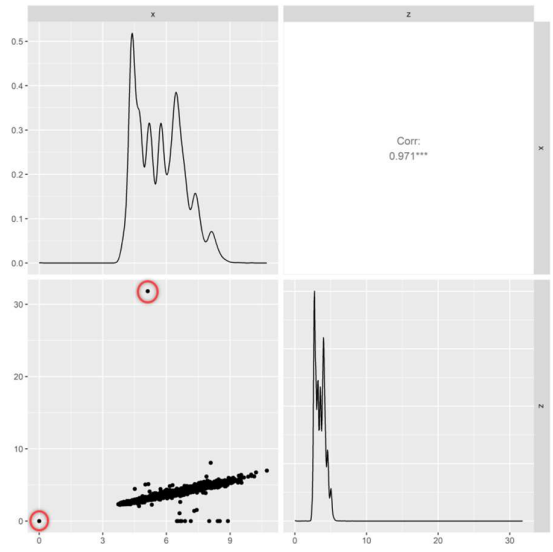
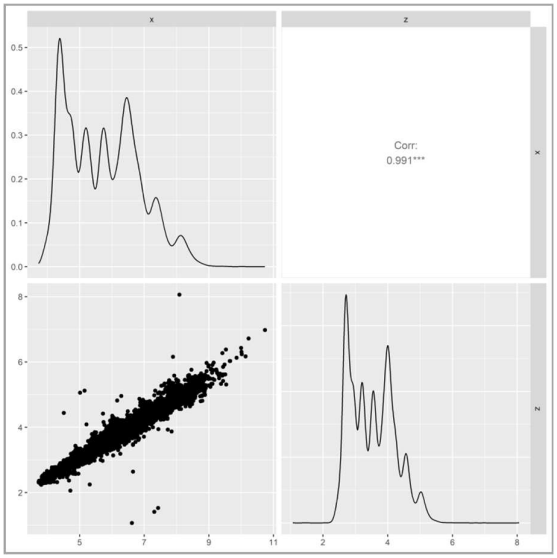


Fig. 24: Pairwise Plot Some Outliers Omitted



3. REFLECTIONS

It is easy to follow instructions without being clear on the main points being covered.

We list some of those points here:

- ***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.
- The first step in almost every dialogue 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.
- Next, we wonder whether you consider Fig. 14 and 15 to be graphs or tables? They have characteristics of both, and the merging of these ideas is one reason the menus in R-Instat are ***Describe*** and ***Model***, rather than the more traditional ***Graphs*** and ***Statistics***.
- Now another statistical point. Earlier in this tutorial we claimed that the diamonds data was well prepared, and it has been used repeatedly by others. So we were able to omit the ***Prepare menu*** and start immediately with the ***Describe menu*** for our analyses.

Fig. 25: Describe Menu

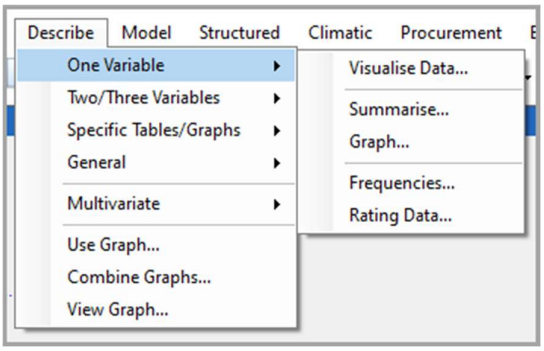
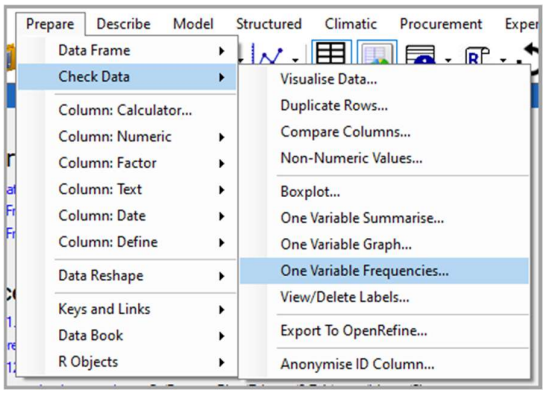


Fig. 26: Prepare menu



- Then the correlation analysis, in Fig. 23, indicated that there were some oddities in the data, that perhaps should have been examined first. We

therefore went back a step to the ***Describe > One Variable > Frequency*** dialogue (Fig. 25) to examine the outliers in more detail. Most of the dialogues in that menu do "double-duty" and are also available in the ***Prepare > Check Data Menu***. Just because the data has already been used extensively by others is not a sufficient reason to omit your own data checking.

- Our main purpose in this tutorial is for users to start learning how to use R via R-Instat. But we are immediately able to combine this, with some statistical ideas. We can be productive in teaching and learning statistics at the same time as we master R-Instat.

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.

5. FEEDBACK AND REPORTING BUGS

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

1. For general feedback 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. 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 if you can be as specific as possible, and detail how to reproduce the bug. Attach data if possible. You may be able to include a capture of the completed dialogue. Ideally, (if this is possible) also paste the R code from the Output window or from the log file.

R-Instat Team, African Data Initiative