Introductory Tutorial: Part 2 A Second Data Set

Introduction

This tutorial guide follows on from Part 1 of the introductory tutorial. We recommend starting with Part 1, although this part is independent of the data and steps from Part 1.

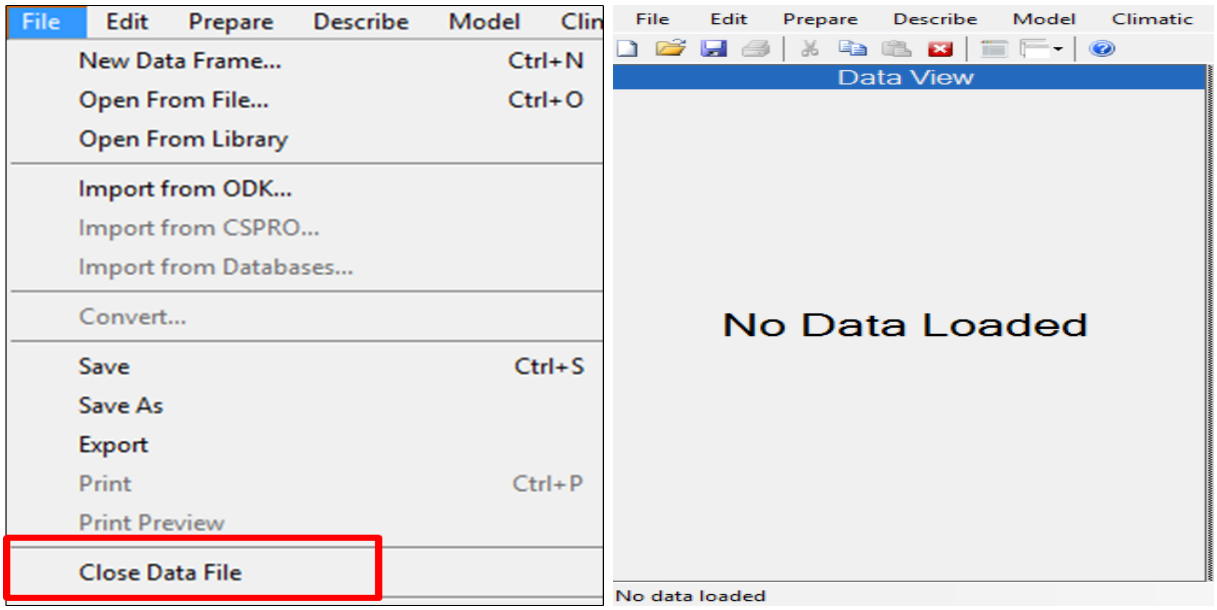
1. The Dodoma data set

This is daily climatic data from Dodoma in Tanzania, from 1935 to 2013. (Footnote: We are very grateful to the Tanzania Met Authority who have given permission for these data to be used for training purposes.)

 If the diamonds data are still in R-Instat then use **File > Close Data File**, Fig. 16.

 You will be asked if you are sure. Respond **Yes**.

Fig. 16. Closing the previous data file **To start again**

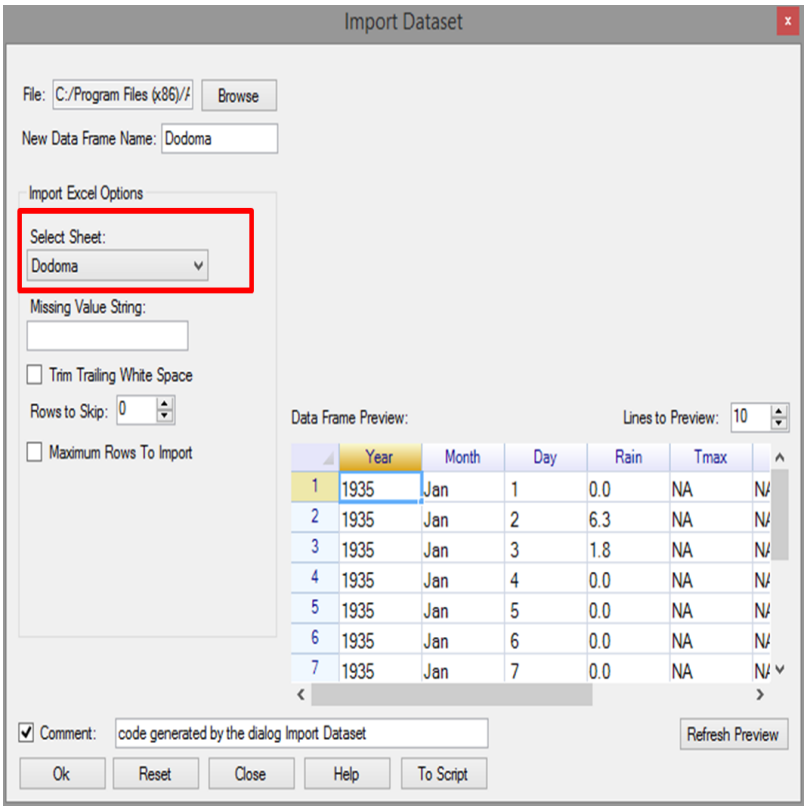


 Use **File > Open from Library**. This time take the option to **Load from Instat Collection** and then press **Browse**.

 Choose **Climatic** and select the Excel file **Climatic_guide_datasets**.

 This Excel file has multiple sheets. Choose the one called **Dodoma**, see Fig. 17

Fig. 17 Opening the Dodoma sheet

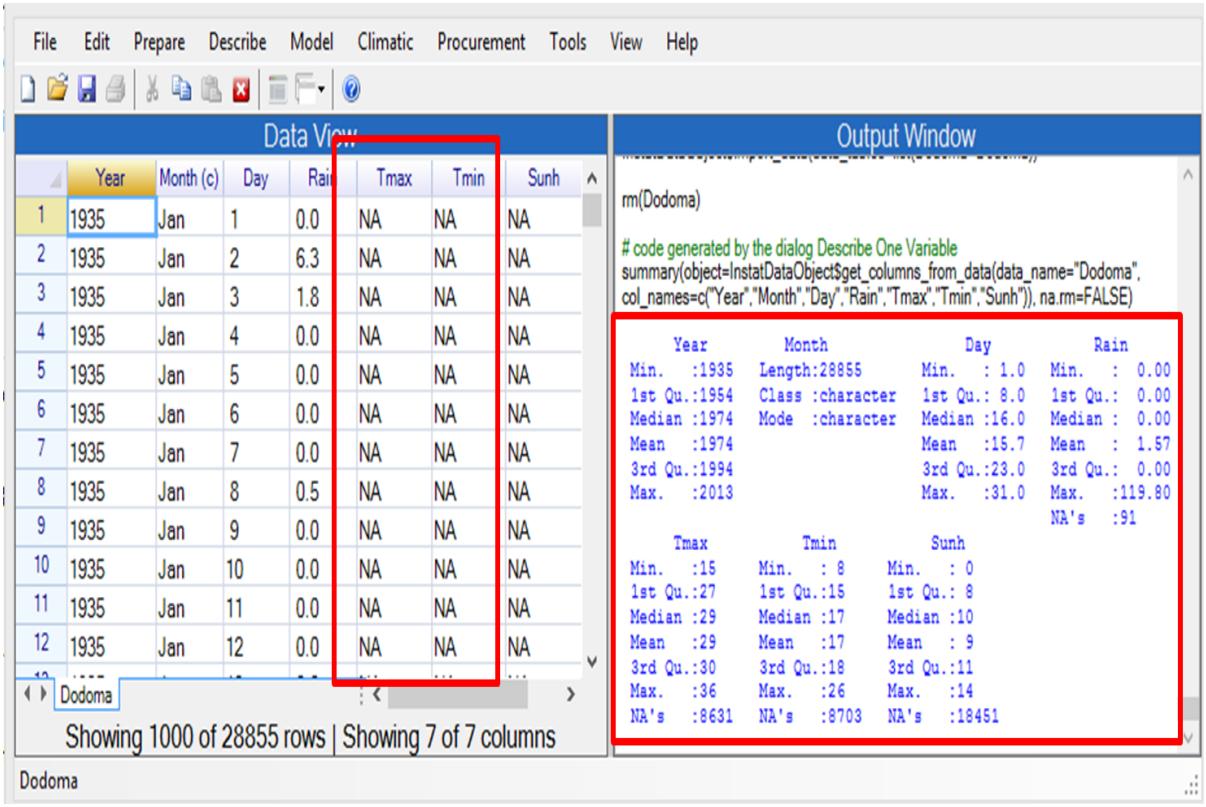


An initial objective is to provide time series graphs for the **annual mean temperatures**, both maximum and minimum . The data are daily, and hence have first to be organised for this analysis. Hence dialogues in the **Prepare menu** will be used, to put the data in the "right shape" for the analysis.

The data are shown in Fig. 18. There are 28,855 observations.

One difference from the diamonds example in Part 1 is that missing values are immediately visible in the data.

Fig. 18 The Dodoma daily data and a summary



 Use the **Describe** > **One Variable** > **Summarise** dialogue again, as in Fig. 12, to produce the summaries also shown in Fig. 18.

The results include the number of missing values, when they exist and over 8 thousand values are missing for the temperature columns. (Hence, as this output was not evident in the similar output in Part 1 (Fig. 12) it follows that the diamonds data did not have any missing values.)

The rainfall data in Fig. 18 are from 1935. Sometimes the stations added temperature records later.

 Use the right-click option on the bottom tab to view the whole data, as shown earlier in Fig. 4.

 Scroll down these data to confirm that the temperatures started from 1958.

This indicates that most of the missing temperature data are because of the later start of measuring these elements.

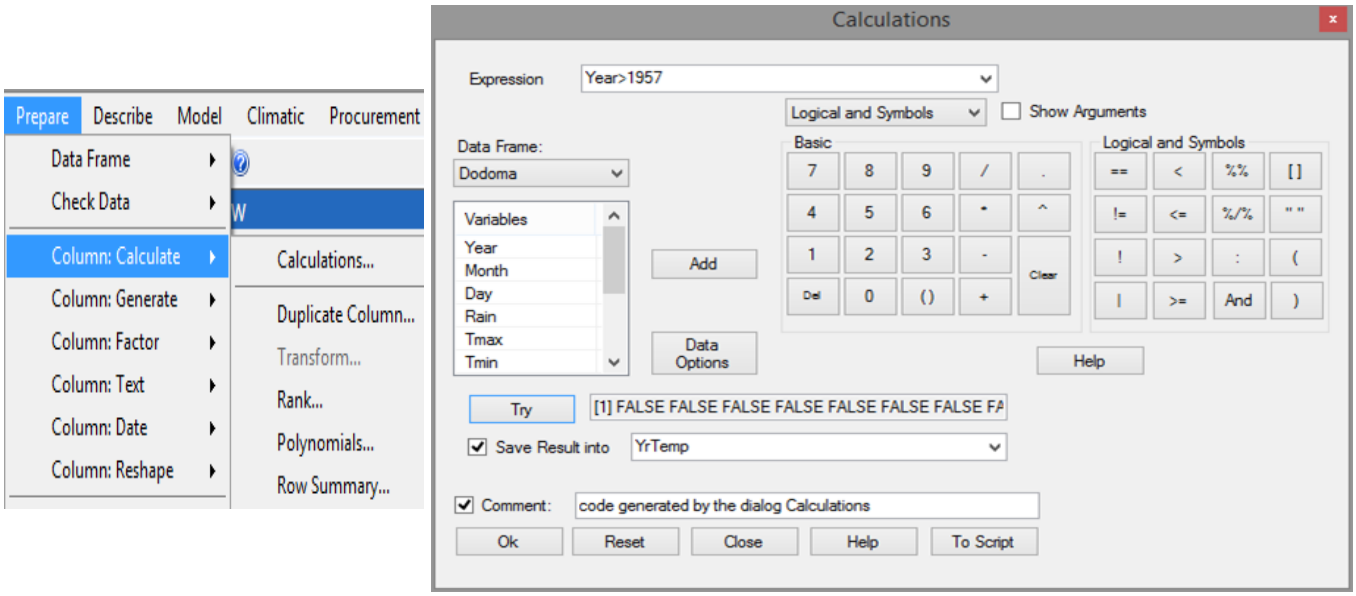
We hope you enjoy, or at least tolerate, the steps below. Often preparing the data for analysis takes most of the time. We have tried to make the Prepare menu in R-Instat as simple to use as possible. There are 5 steps to go through even for the simple tasks here. But there is a "silver lining" at the end, as we explain in Section 4!

2. Preparing the data

Often the preparation of the data includes calculating further columns.

 Open the **Prepare** > **Calculations** > **Calculate** dialogue as shown in Fig. 19.

Fig. 19. The prepare menu With the calculate dialogue



This is designed to be a column calculator. It has multiple keyboards.

 Click on the control that currently says **Basic** and choose **Logical and Symbols**. An additional keyboard opens as shown in Fig. 19.

 Double-click on the Year column, (or click and press Add) to put it into the formula field at the top of the dialogue.

 Complete the formula by adding **> 1957**, so it reads **Year > 1957**, see Fig. 19.

 Click on the **Try** button and it should give the result **FALSE, FALSE, FALSE...** as in Fig. 19, because the first rows of data are from 1935 - hence not more than 1957!

 Give a name for the new column to save the results, like **YrTemp**. Then press **OK**.

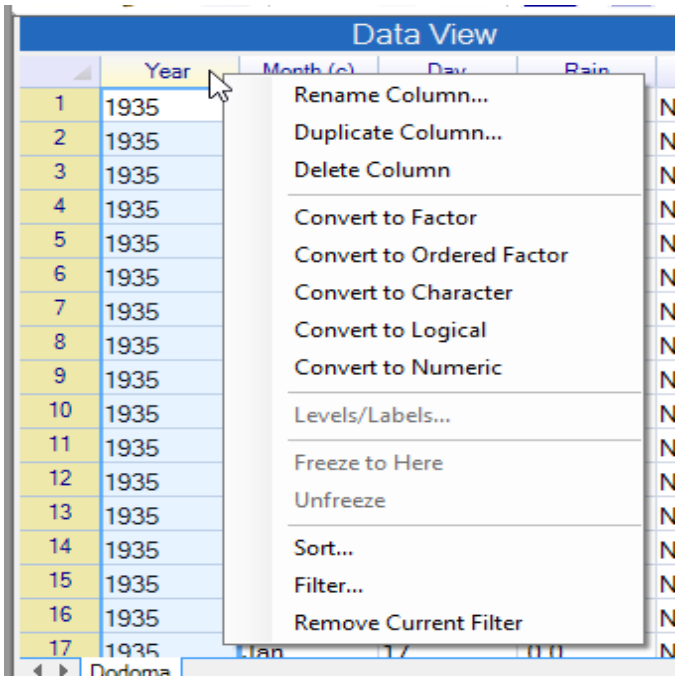
This should produce a new column of data.

The next step is to apply a **filter**, so the data for analysis only start in 1958, i.e. when the new column just produced is TRUE. Many common tasks from the Prepare menu are quickly accessible through a special **right-click menu** which is shown in Fig. 20.

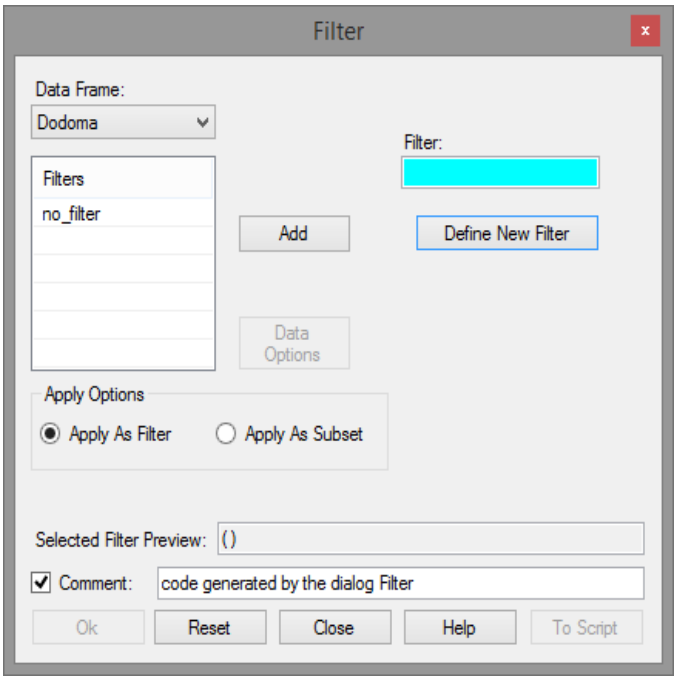
 Put the cursor in the top row (with the names) and **right-click**, Fig. 20.

 Choose the **Filter dialogue** from this menu, Fig 20.

Fig. 20. The right-click menu



To choose a filter

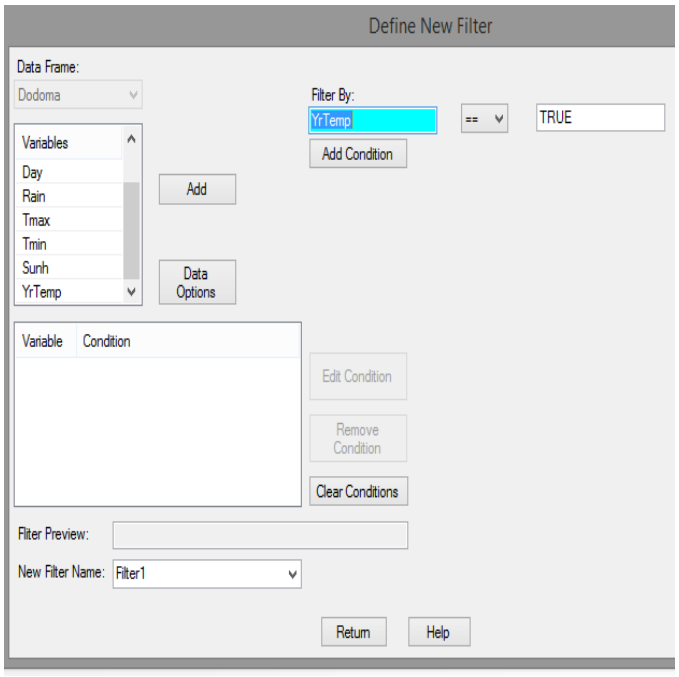


- Click on the button in Fig. 20 to **Define New Filter**.

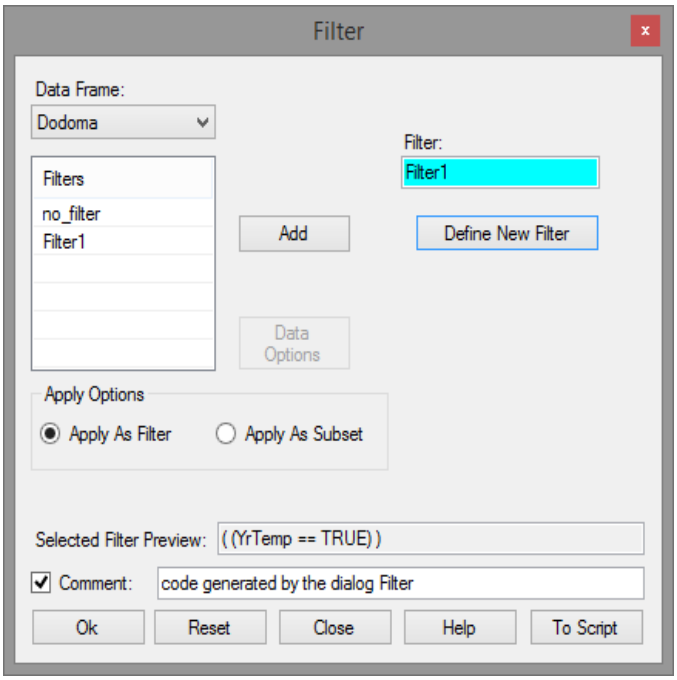
- In the sub-dialogue, choose the YrTemp column. Complete the condition so it reads **YrTemp == TRUE**

(Note the == is not a mistake, and the word **TRUE** must be in capital letters, Fig. 21)

Fig. 21 Specifying the filter



And then applying it



- Press the button to **Add Condition**, Fig. 21 and then press **Return**.

- On the main filter dialogue, Fig. 21, press **OK** to apply the filter.

Note the first column, with the row numbers, is now in red and the first one is row 8402, i.e. 1

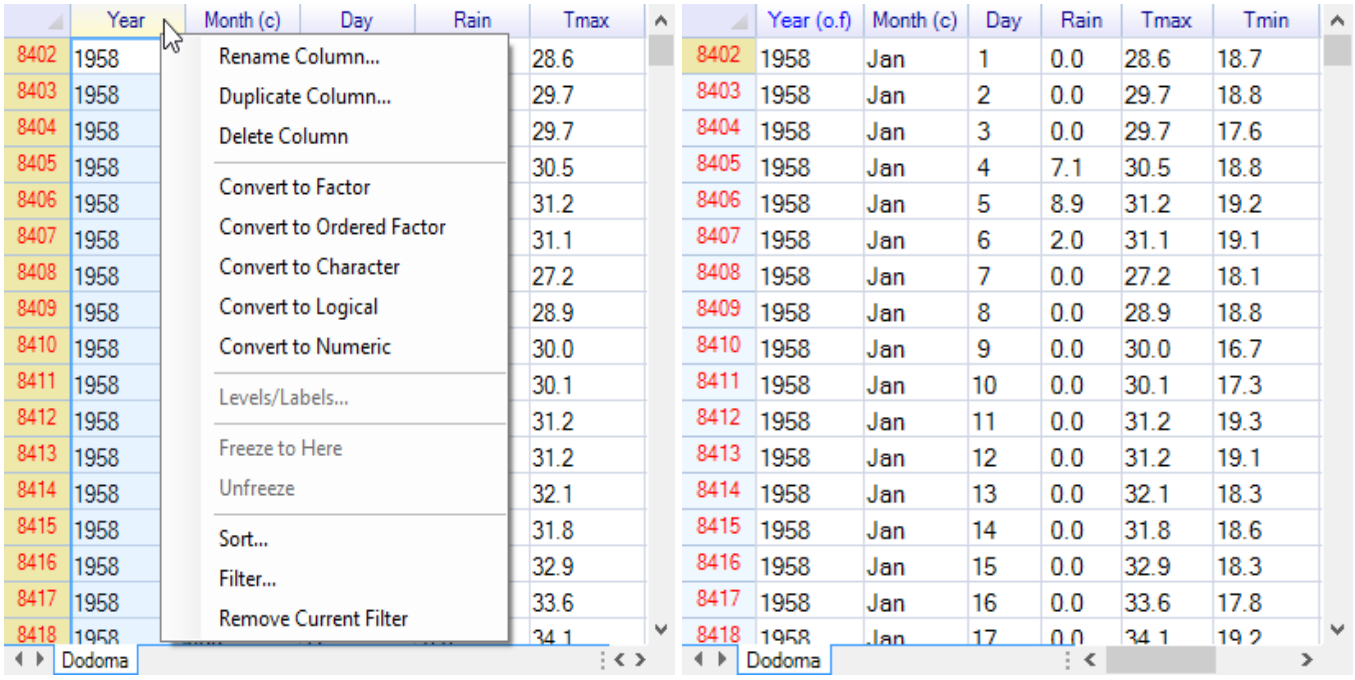
January 1958.

The third preparatory step is to **change the Year column**, which is numeric, into a category, or **factor** type of column.

 Go to the **Year** column and to the top (name) row. **Right-Click**, Fig. 22.

 Click on **Convert to Ordered Factor**.

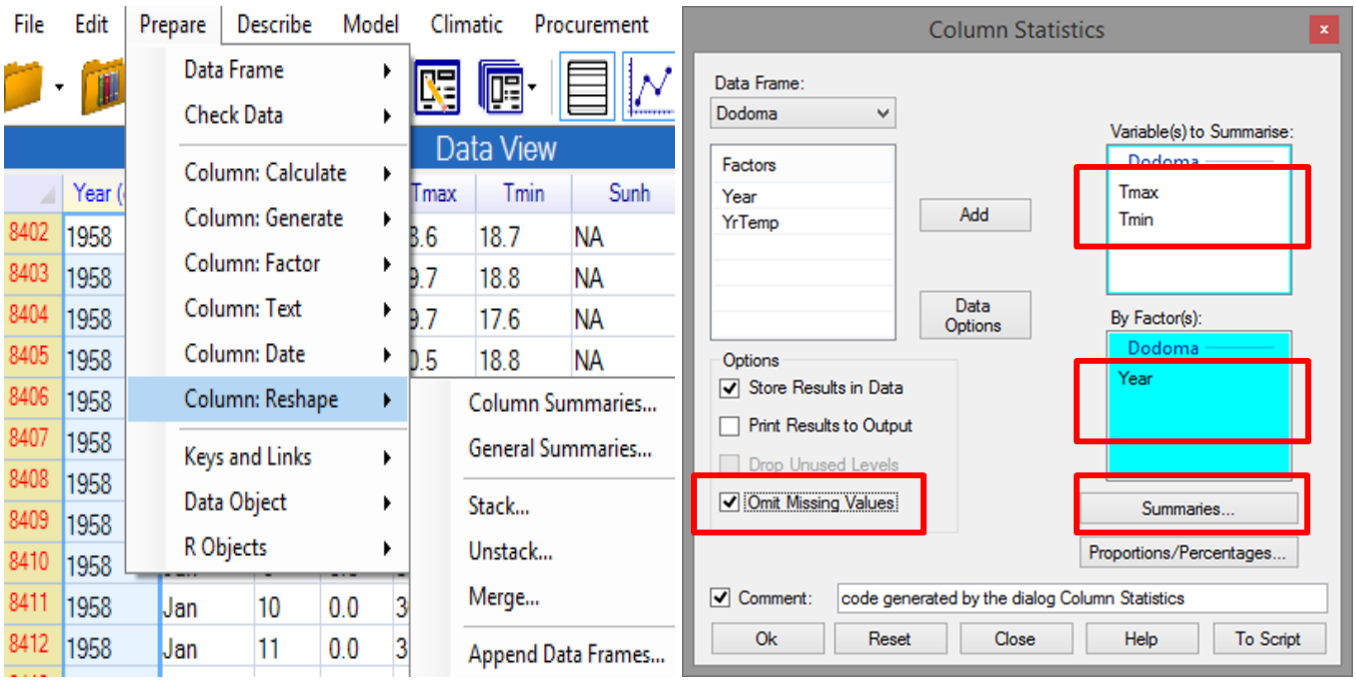
Fig. 22. Converting the Year column to an ordered factor **The resulting data**



The daily data are now ready to be summarised to produce the yearly means.

 Open the **Prepare > Column: Reshape > Column Summaries** dialogue, Fig 23.

Fig. 23. Menu for Column Summaries **With the resulting dialogue**



 Complete the dialogue as shown in Fig. 23, i.e. **Tmin and Tmax** into the main receiver, **Year** into the other receiver, and the option ticked to **Omit Missing Values**.

 Then press the **Summaries button** to move to the sub-dialogue, Fig. 24.

 Complete the sub-dialogue as shown in Fig 24, i.e. with only two summaries for the **N Not Missing** and the **Mean**. Then press **Return**.

 Press **OK** to produce the summaries, Fig. 24.

Fig. 24. Summaries sub-dialogue **With the resulting data**

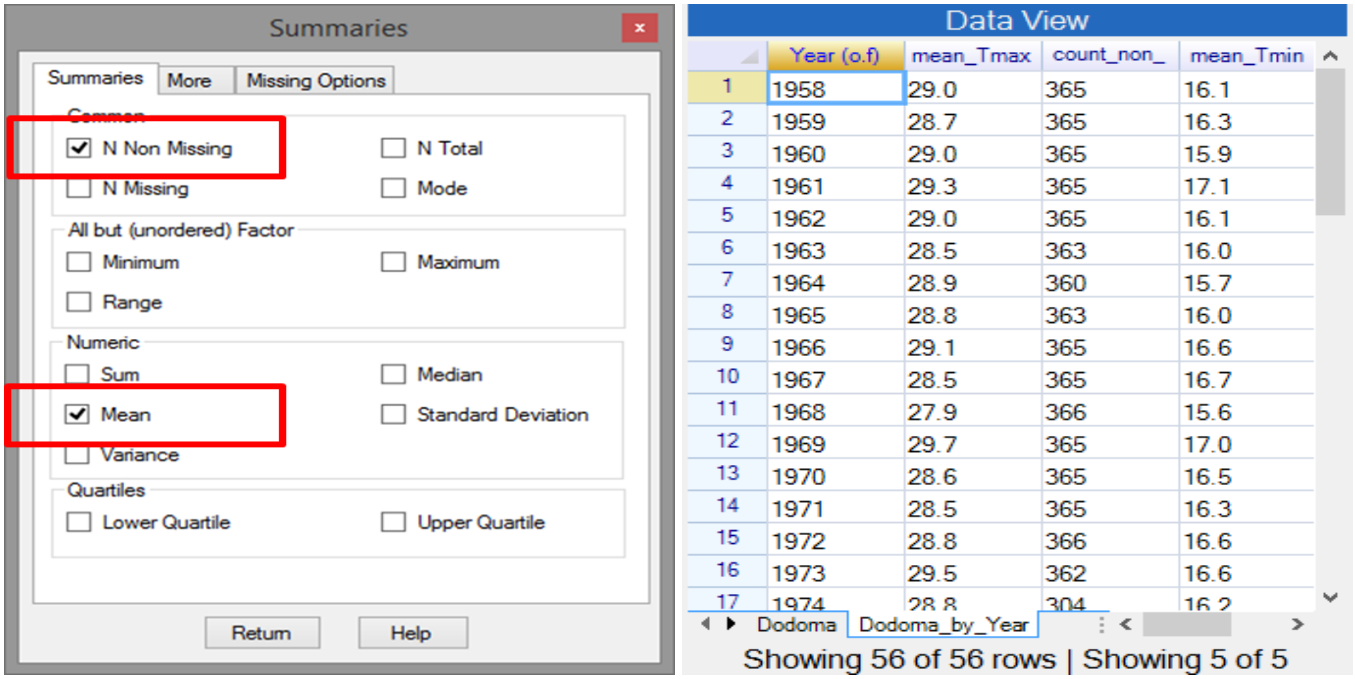


Fig. 24 also shows we now have **2 data frames**, one at the daily level and the other with the annual summaries. This second data frame is needed for the graphs.

3. Producing the graphs

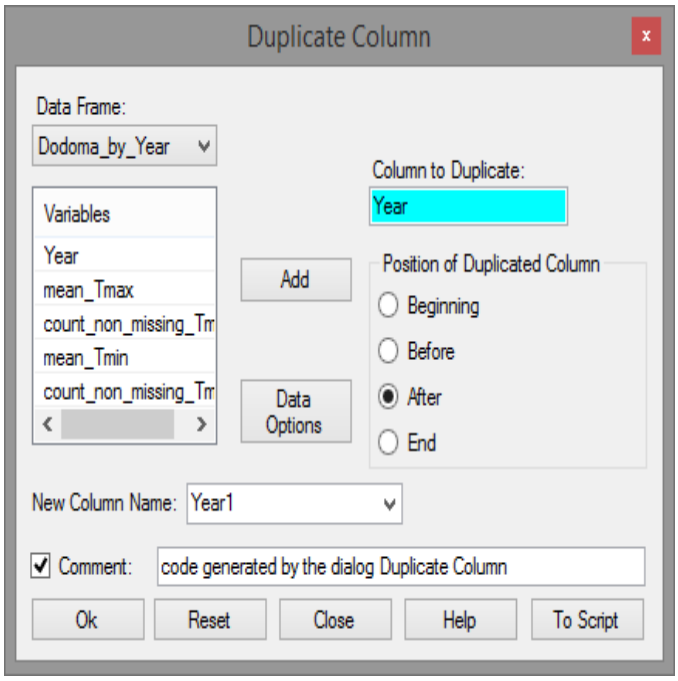
We have one final small preparatory step to do first. This is because the Year column in the Summary data is a factor column. For the graphs we need it to be numeric again. It is often convenient to have both!

 Use **Prepare** > **Calculate** > **Duplicate Column** (or right click and choose the appropriate item.)

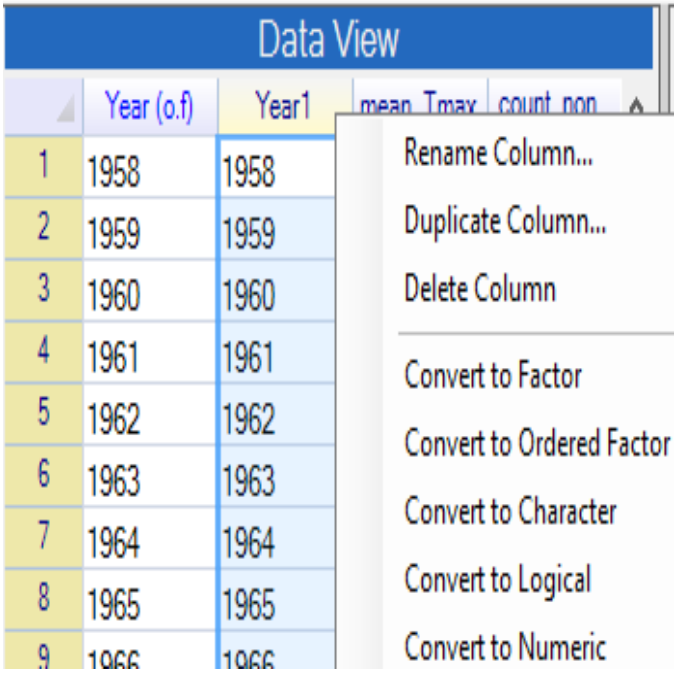
 Complete the dialogue as shown in Fig. 25. Press **OK** to produce another column called **Year1**.

 Right-click on the **Year1** name and make the column **numeric** Fig. 25.

Fig. 25. Duplicating a column



Making the resulting column numeric

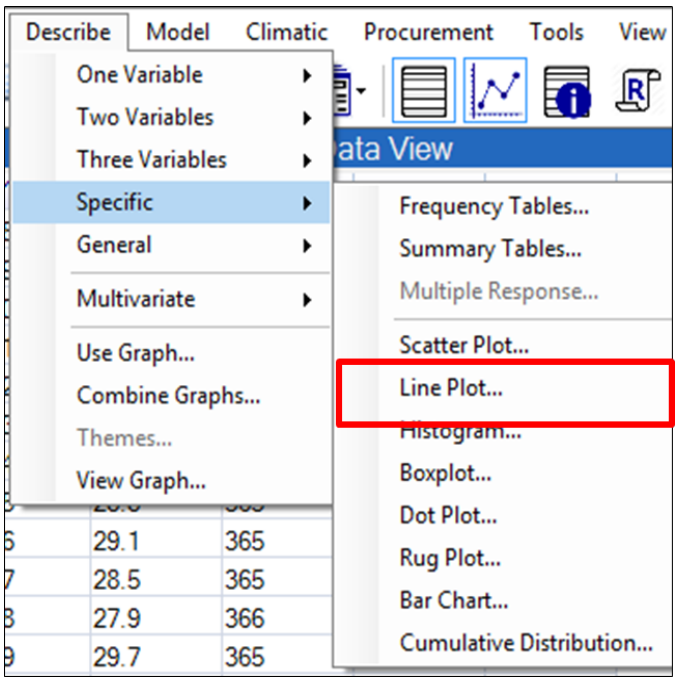


At last we are ready to produce the graphs.

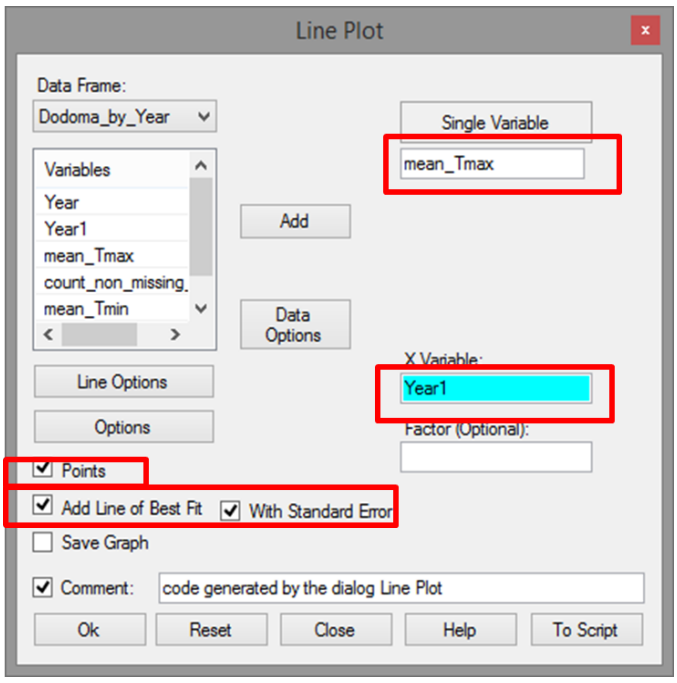
 Use **Describe** > **Specific** > **Line Plot**, Fig. 26.

 Complete the dialogue as shown in Fig. 26 for the **mean_Tmax**. Press **OK**.

Fig. 26. The line plot menu



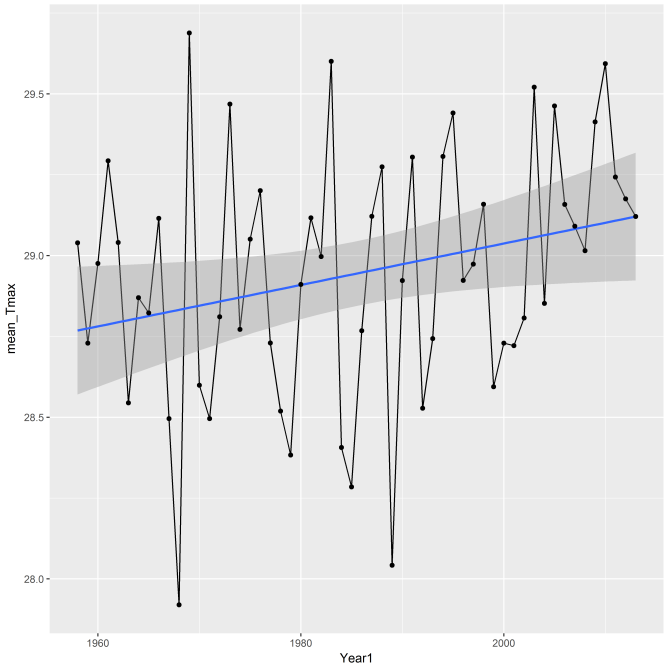
And the dialogue



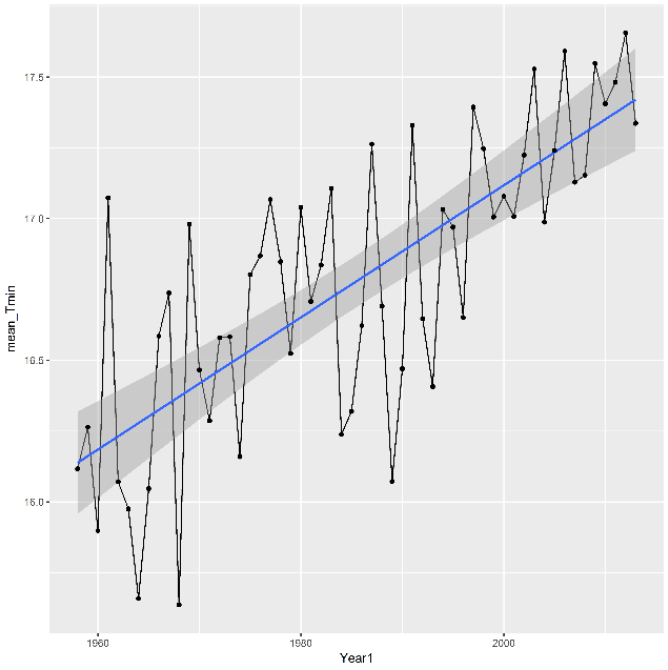
The resulting graph is shown in Fig. 27.

 Return to the Line Plot dialogue and swap **mean_Tmin** for **mean_Tmax**. Press **OK** to give the second graph also shown in Fig. 27

Fig. 27. The graph for Tmax



And for Tmin



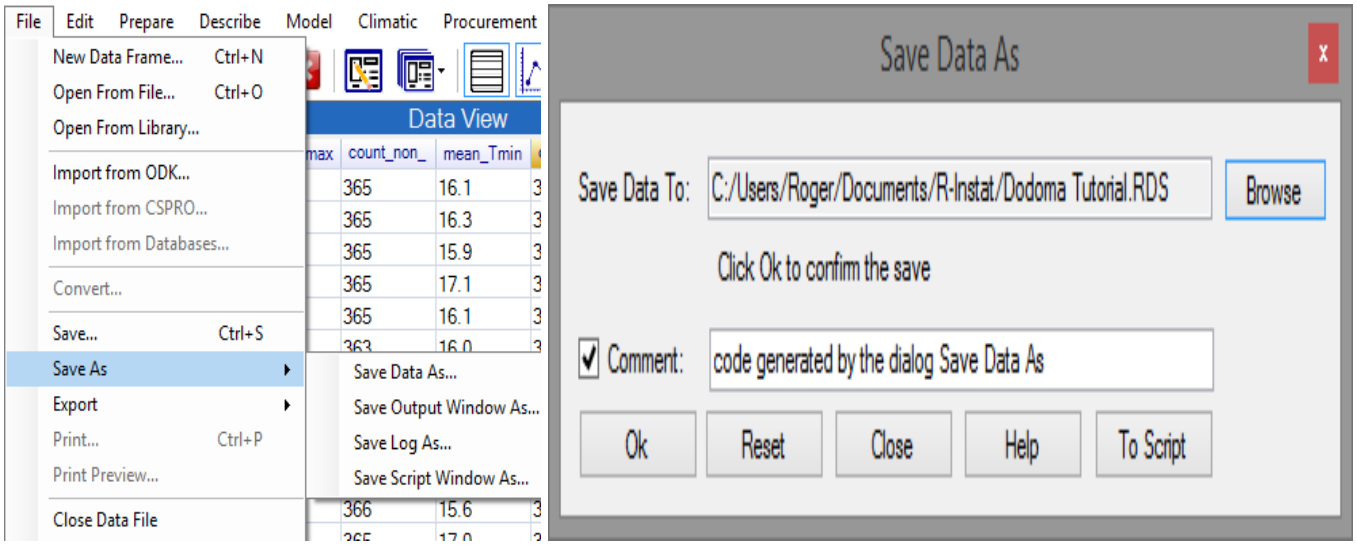
4. Saving the data

Before using a different data set save these data, so you could resume later.

 Use the **File ? Save As** dialog, Fig. 28. Choose the option **Save Data As**.

 Press on **Browse** in the dialogue, Fig. 28. Choose a suitable directory and name. Press **OK** when you return to the Save Data dialogue.

Fig. 28. Saving the data set



The RDS extension is added, to signify it is saved as an R data file. This is the "silver lining" we mentioned in Section 1. If done well, the data only have to be organised once. Then the resulting file, with the two data frames, can be opened in the future, and the analysis can be continued.

5. Next steps

There are more analyses that can be explored with this data in R-Instat and we encourage you now to try. The next part of the tutorial focuses on working with labelled data.

6. 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 (at) 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