Preparing climatic data for analysis with R-Instat -

12 July 2018

Table of Contents

1) Introduction

R-Instat is designed as a general statistics package. All the calculations are made through the statistical system called R. In addition, there is a special climatic menu.

In the future we plan for the climatic menu to facilitate the analysis of climatic data at any scale, e.g. from an automatic station. Currently the methods are particularly designed for daily data. This guide uses an example of daily data for 2 stations in Guinee (Conakry) each of which was supplied with 4 elements. These data are in the R-Instat library and hence the examples in this guide can be followed by users who wish to do so.

2) Acknowledgement

We gratefully acknowledge the permission of the Guinee Met Service for their permission to use their data in preparing this guide, and to allow their data to be added to the R-Instat library.

3) Getting the data into shape

We first open the data in roughly the form it was originally presented. This is in 2 Excel files and hence we start by showing the data in Excel, rather than in R-Instat, Fig. 1.

Fig. 1 The data for one station in Excel

	A	B	C	D	E	F
1		TMIN journalière				
2	Eg gh id	Eg el abbrevi	Year	Month	Day	Value
3	17KKAN1S	TMIN	1950	01	01	13
4	17KKAN1S	TMIN	1950	01	02	15.6
5	17KKAN1S	TMIN	1950	01	03	18
6	17KKAN1S	TMIN	1950	01	04	19.7
7	17KKAN1S	TMIN	1950	01	05	14
8	17KKAN1S	TMIN	1950	01	06	12.4
9	17KKAN1S	TMIN	1950	01	07	11.9
10	17KKAN1S	TMIN	1950	01	08	18.4
11	17KKAN1S	TMIN	1950	01	09	13.2

Tmin | Tmax | Rain | RelHum

Fig 2 The shape used by R-Instat

Data View

	Station (c)	Year	Month	Day	Rain	RelHum	Tmax	Tmin
1	Kankan	1950	01	01	0.0	NA	35.8	13.0
2	Kankan	1950	01	02	0.0	NA	34.8	15.6
3	Kankan	1950	01	03	0.0	NA	34.7	18.0
4	Kankan	1950	01	04	0.0	NA	33.9	19.7
5	Kankan	1950	01	05	0.0	NA	33.8	14.0
6	Kankan	1950	01	06	0.0	NA	31.2	12.4
7	Kankan	1950	01	07	0.0	NA	33.0	11.9
8	Kankan	1950	01	08	0.0	NA	34.4	18.4
9	Kankan	1950	01	09	0.0	NA	34.3	13.2
10	Kankan	1950	01	10	0.0	NA	33.5	12.5
11	Kankan	1950	01	11	0.0	NA	33.3	12.6
12	Kankan	1950	01	12	0.0	NA	33.5	12.5
13	Kankan	1950	01	13	0.0	NA	32.0	11.0
14	Kankan	1950	01	14	0.0	NA	29.2	12.3

Kankan | Koundara | Guinee2

Showing 1000 of 40475 rows | Showing 8 of 8 columns

The data are in the right "shape" for R-Instat, i.e. one row of data for each day. The data for this station start in 1950 and continue to 2016 or 2017.

This is not always the case and in Appendix 1 we consider how to transform data that start in different "shapes".

If your data are already in the right "shape", as shown in Fig. 2 (and also Fig. 22) then you could jump to Section 4 of this guide. This shape has the 4 elements and both stations in the same data frame. The elements in Fig. 2 are in successive columns. The stations are one below the other, with the first column in Fig. 2 giving the station name.

There are also 4 sheets in the Excel file, Fig. 1, with a different element on each sheet.

If the analysis is only for a single element, and for just this one station, then these data can be imported into R-Instat as shown below, and then continuing with Section 4. However, many analyses benefit from the data in these 4 sheets being merged into a single file. They can also be combined with those from the second station. That is what we show here.

Our aim is for the four elements, for both stations, to be in a single sheet or data frame.

****Go into R-Instat***.

Use ***File > Open from Library* to open the data file, see Fig. 3.

(That is because these data are in the Instat library. For your own data use ***File > Open*** instead and look for the file.)

Fig. 3 File > Open from Library

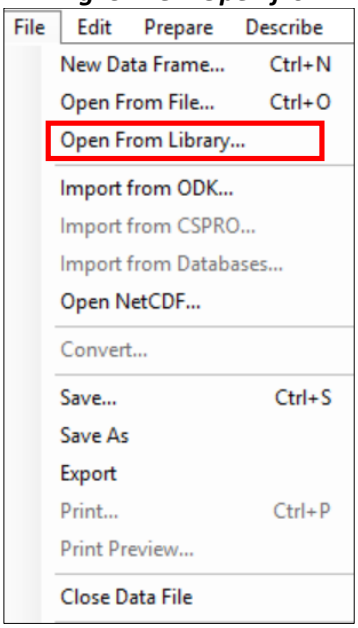
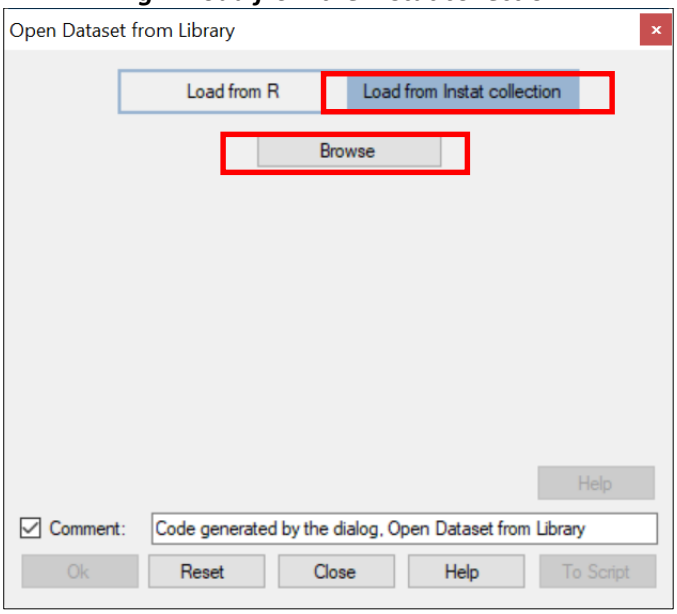


Fig. 4 Load from the Instat collection



Click on ***Load from Instat collection*, Fig. 5.

****Click Browse,*** then choose ***Climatic and then Guinee***.

Fig. 5 Choose the Excel file

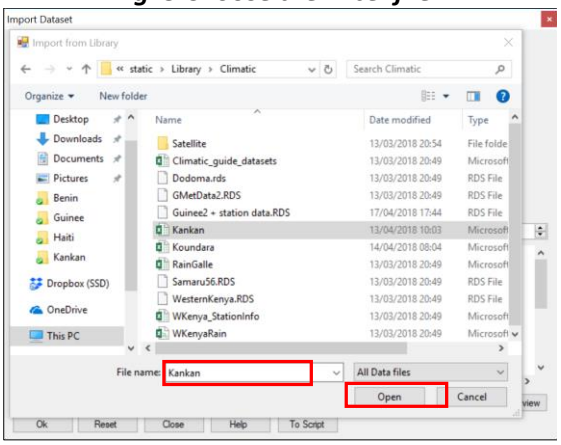
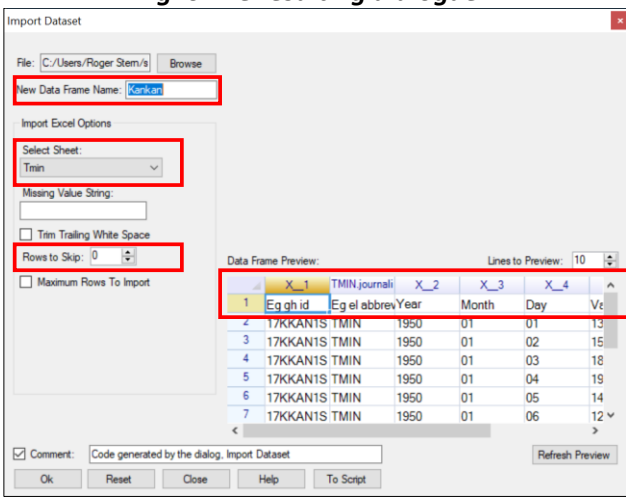


Fig. 6 The resulting dialogue



Choose the file called ***Kankan.xlsx* and ***click on Open***, Fig, 5.

** ***Examine the dialogue***, Fig. 6. Do NOT yet click OK. The current Excel sheet is not ready to import.

One issue is clearer from another look at the data in Excel in Fig. 1. The first row is a heading, and the variable names are in row 2 of the sheet.

Change the ***Rows to Skip to 1*, Fig. 6.

** C***hange the name*** of the resulting data frame ***to Tmin.*** The results now look as in Fig. 7.

Fig. 7 The data ready to import

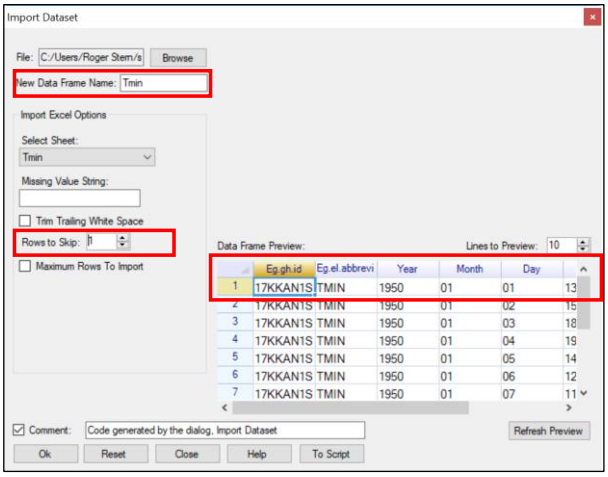
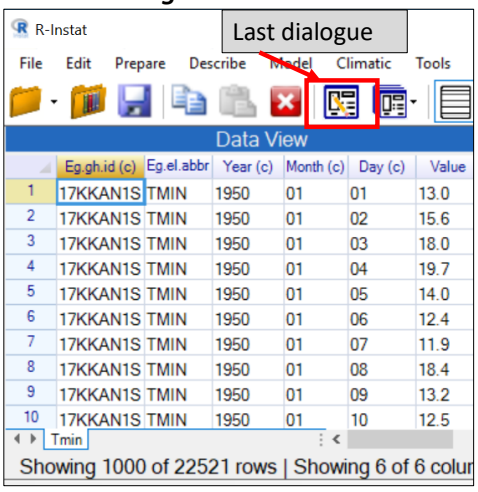


Fig. 8 Data in R-Instat



****Click OK*** to import the data into R-Instat, Fig. 8.

The bottom of Fig. 8 indicates there are 22,521 days, i.e. rows of data, of which only 1000 are shown. The grid in R-Instat is just a window showing part of the data. They are stored in R.

** Click on the ***Last Dialogue*** button, see Fig. 8. This returns to Fig. 7.

** ***Change the Select Sheet*** to the second Excel sheet, i.e***. Tmax***.

** ***Change the name to Tmax***

****Press OK.***

You now have 2 data frames in R-Instat. The Tmax data frame has 22215 rows of data.

** ***Repeat twice more*** to import ***Rain and RelHum*** into R-Instat. You should now have 4 sheets or data frames in R as shown in Fig. 9.

There are 24080 rows of rainfall data and 11,045 rows for relative humidity.

Fig. 9 The 4 data frames in R-Instat

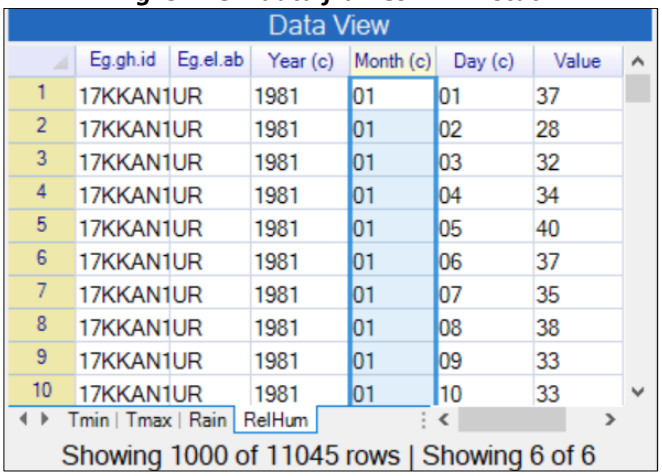
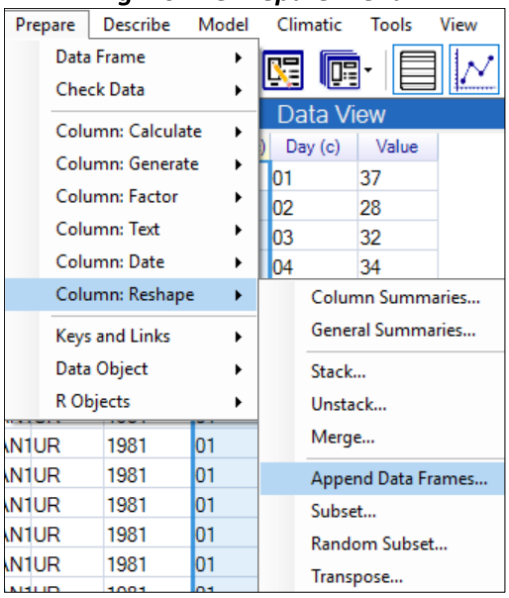


Fig. 10 The Prepare menu



The next task is to merge the data for the different elements into a single data frame. This is done in 2 stages.

As we are preparing the data, this uses the Prepare menu, Fig. 10

** In the ***Prepare menu,*** choose ***Column: Reshape*** and then ***Append Data Frames***, Fig. 10.

Fig. 11 The Append dialogue

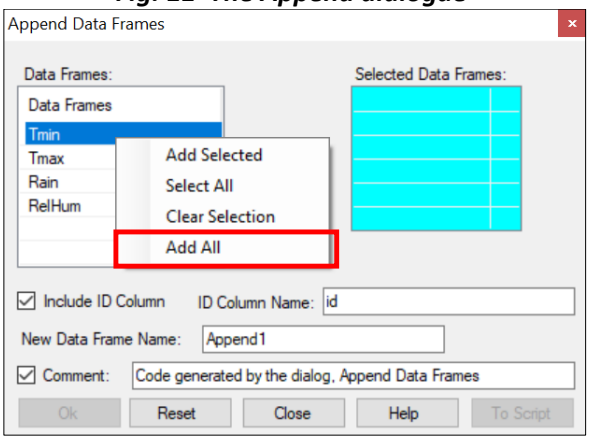
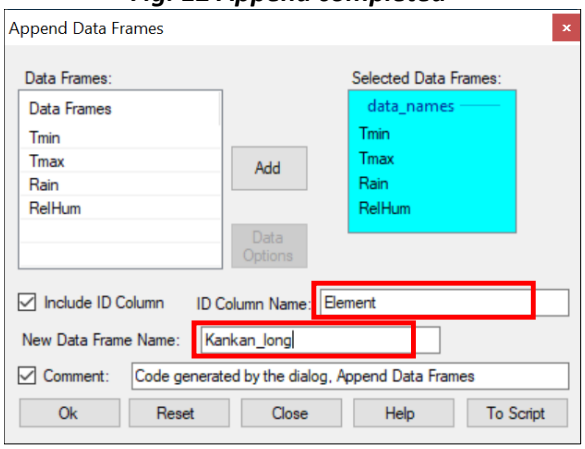


Fig. 12 Append completed



In the resulting dialogue you want to use all the data frames.

** ***Right-click*** in the data selector, see Fig. 11. Then choose ***Add All***.

** ***Change the ID column name*** to ***Element*** and the ***Data Frame name*** to ***Kankan_long***. Fig. 12.

** ***Press OK***,

This has produced a new data frame with 79,861 rows of data.

The first column, called Element is a text (character) type of column. It must be converted into a category type, which is called a Factor column in R.

** ***Right-click on the name Element***. This gives the pull-down menu shown in Fig. 13.

** Choose ***Convert to Factor***. An (f) now appears after the column name.

Fig 13. The right-click menu

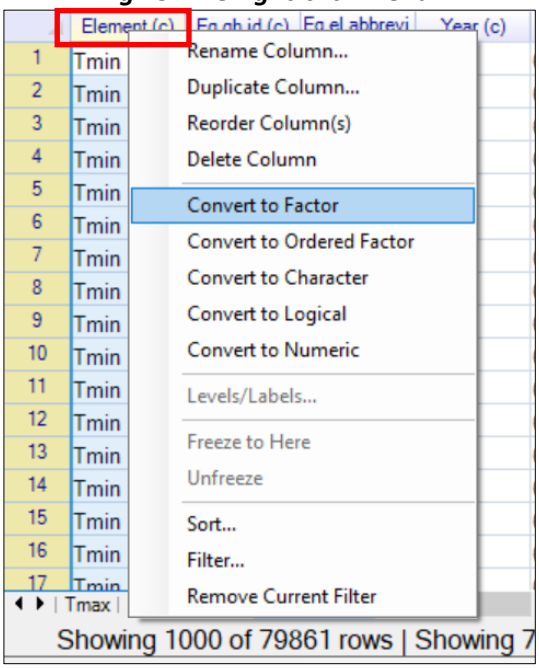
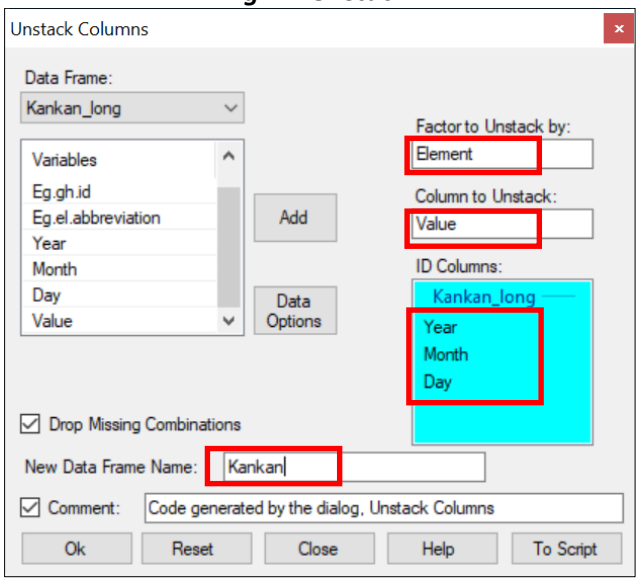


Fig. 14 Unstack



** Use ***Prepare > Column: Reshape > Unstack***, see Fig. 9 again.

Complete the unstack dialogue as follows, see Fig. 14.

** The ***factor*** is the ***Element column***, while the ***Column to Unstack*** is ***Value***.

** There are 3 ***ID columns***, namely ***Year, Month, Day*** and the ***New Data Frame*** is called ***Kankan***.

** Press ***Ok***.

Fig. 15 Right-click on the data frame

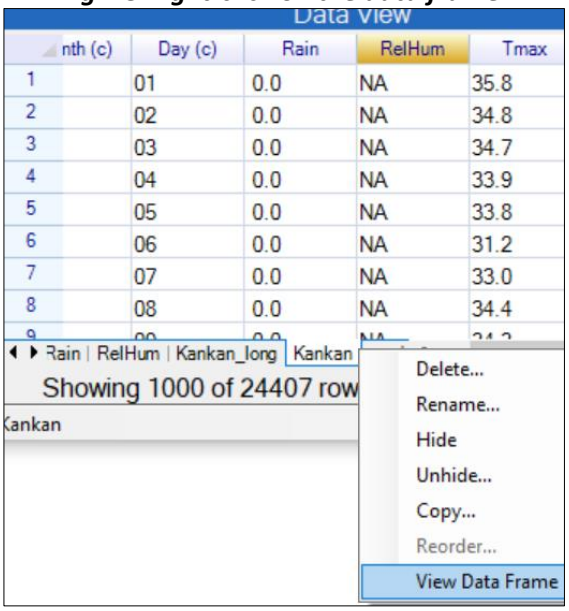
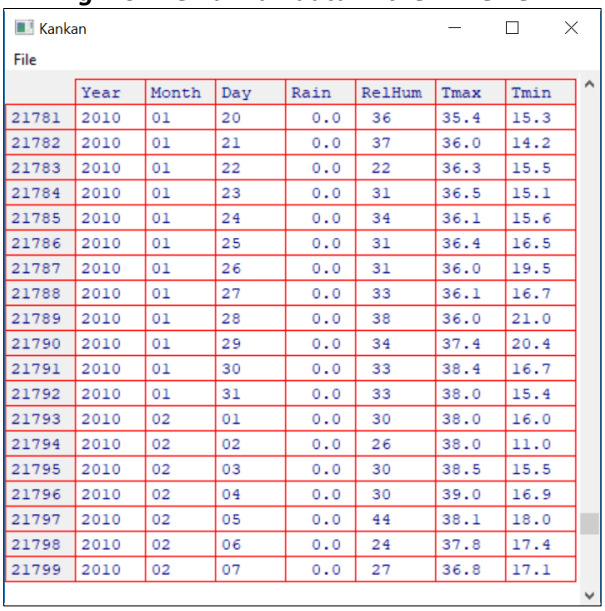


Fig. 16 The Kankan data in the R-viewer



Kankan

File

	Year	Month	Day	Rain	RelHum	Tmax	Tmin
21781	2010	01	20	0.0	36	35.4	15.3
21782	2010	01	21	0.0	37	36.0	14.2
21783	2010	01	22	0.0	22	36.3	15.5
21784	2010	01	23	0.0	31	36.5	15.1
21785	2010	01	24	0.0	34	36.1	15.6
21786	2010	01	25	0.0	31	36.4	16.5
21787	2010	01	26	0.0	31	36.0	19.5
21788	2010	01	27	0.0	33	36.1	16.7
21789	2010	01	28	0.0	38	36.0	21.0
21790	2010	01	29	0.0	34	37.4	20.4
21791	2010	01	30	0.0	33	38.4	16.7
21792	2010	01	31	0.0	33	38.0	15.4
21793	2010	02	01	0.0	30	38.0	16.0
21794	2010	02	02	0.0	26	38.0	11.0
21795	2010	02	03	0.0	30	38.5	15.5
21796	2010	02	04	0.0	30	39.0	16.9
21797	2010	02	05	0.0	44	38.1	18.0
21798	2010	02	06	0.0	24	37.8	17.4
21799	2010	02	07	0.0	27	36.8	17.1

The grid, or spreadsheet, in R-Instat is just a window showing part of the data.

** ***Right-click*** on the name ***Kankan***, Fig. 15, and choose ***View Data Frame.***

** ***Scroll down*** the data in the R-viewer, Fig. 16, to see the complete data.

Finally, for these data, we now delete the other data frames. The only one needed is Kankan.

** Choose the ***Kankan_long*** data frame.

** ***Right-click at the bottom*** of the window, see Fig. 15 again, and choose ***Delete,*** Fig. 17.

***** Add the other data frames – except Kankan, of course -*** see Fig. 17!

Fig. 17 Deleting unwanted data frames

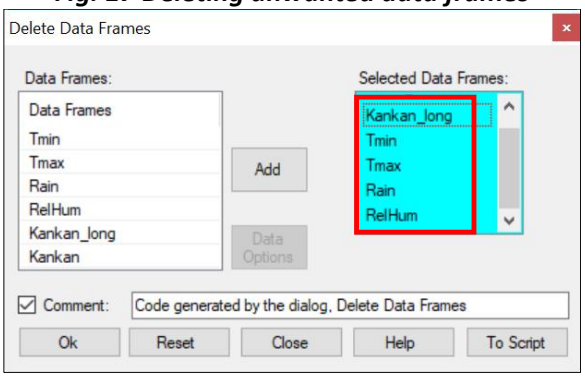
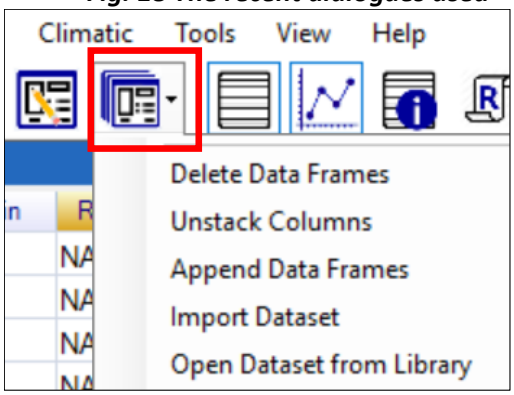


Fig. 18 The recent dialogues used



*****Press OK***.

The dialogues used so far are now also easily available, through the toolbar.

** ***Click on the icon,*** Fig. 18, to see what has been used.

** In the list – Fig. 18 – choose ***Import Dataset***.

** In the resulting dialogue, Fig. 19, use ***Browse*** and choose the second station, ***Koundara.***

Fig. 19 A second data set

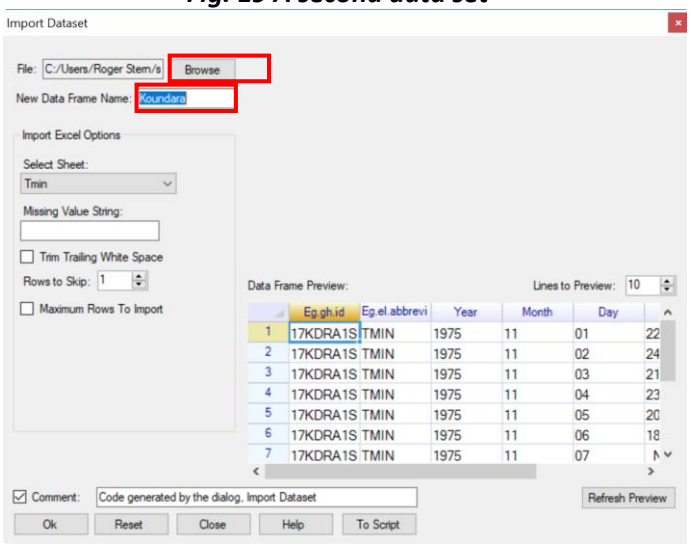
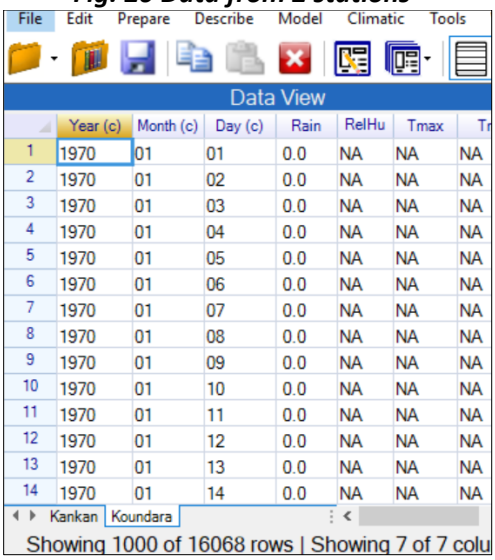


Fig. 20 Data from 2 stations



** C***hange the name*** of the resulting data frame ***to Tmin. Click Ok***.

** Now ***repeat the steps*** from Fig. 8, page 3, to Fig. 17, page 6, for the ***Koundara*** data.

The results, with the 2 data frames are in Fig. 20

** Use ***Prepare > Column: Reshape > Append*** again, see Fig 21 to put the data into a single data frame.

Fig. 21 Appending the data for the 2 stations

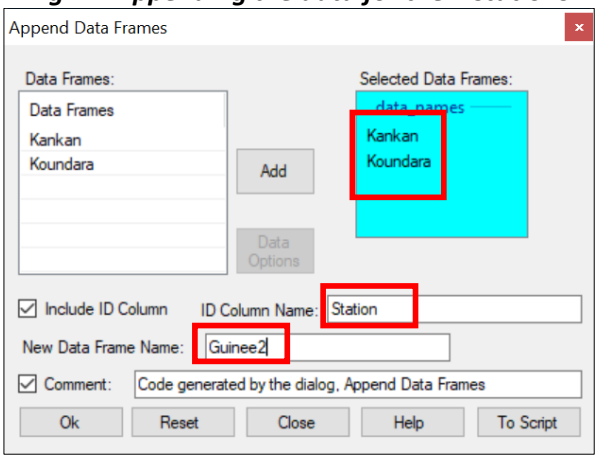
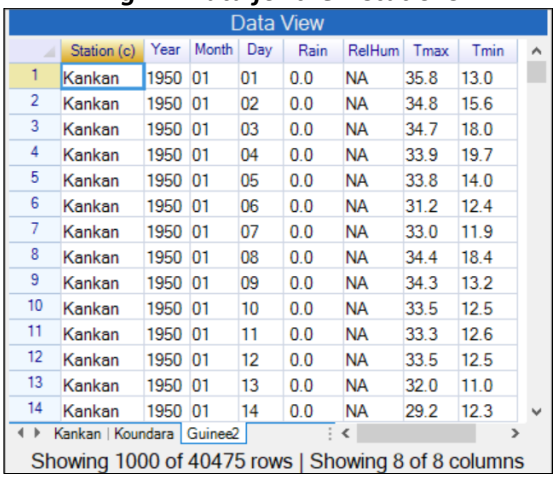


Fig. 22 Data for the 2 stations



This stage is now over. The data for both stations, and with all four elements, are in a single data frame, Fig. 22. There are 40475 rows (days) of data.

To complete the initial task the resulting files are now saved.

** Choose ***File > Export > Export Dataset***, Fig. 23.

Fig. 23 Export and save the data

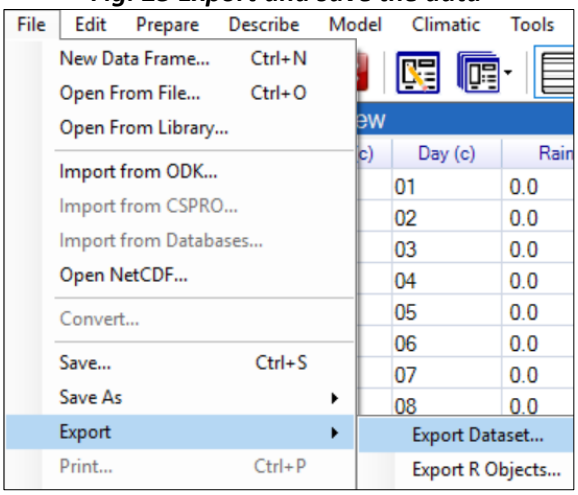
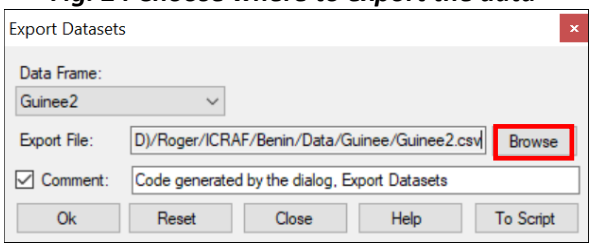


Fig. 24 Choose where to export the data



** Click on ***Browse***, Fig. 24, and choose where to save the exported file.

** After choosing the file name, you return to Fig. 24. ***Click OK***. The file is not saved until ***OK*** is clicked

By default, it is a csv file, which can easily be read into Excel.

** ***Right click*** on the ***Kankan*** name at the bottom of the data frame.

** Select ***Delete***, Fig. 25, to delete the individual station data.

** Choose ***File > Save As > Save Data As***, Fig. 26, to save a file of type RDS for reading back into R-Instat.

Fig. 25. Delete the extra data frames

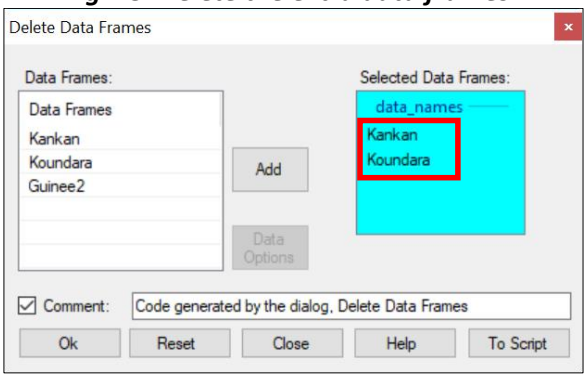
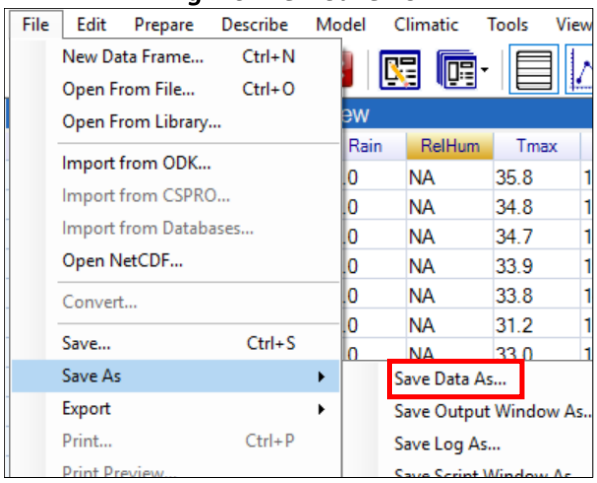


Fig. 26 File > Save As



Review

This first stage has organised the data, ready for climatic analyses with R-Instat. The main task has been to reshape the data into the form used by R-Instat.

The details of this stage depend on the original "shape" of the data. Other common starting points are considered in Appendix 1.

We used the File menu and then the Prepare menu in R-Instat. Within the Prepare menu we particularly used the sub-menu Prepare > Column: > Reshape and then used the Append and Unstack dialogues. Other data formats use additional options from this sub-menu, particularly Stack and Merge.

4) Adding a date column

** If you are continuing from the section above, that is fine. Continue with the data.

(** Otherwise start by loading the data you saved above. Or go again to ***File > Open from Library > Load from Instat Collection > Browse > Climatic > Guinee*** and choose the file called ***Guinee2.csv.)***

Fig. 1 Convert Station to a factor column

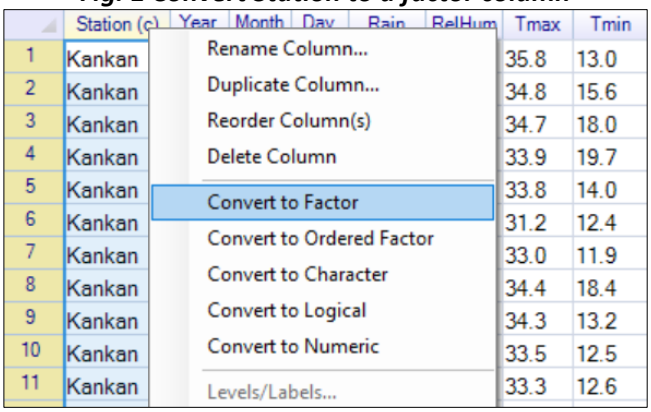
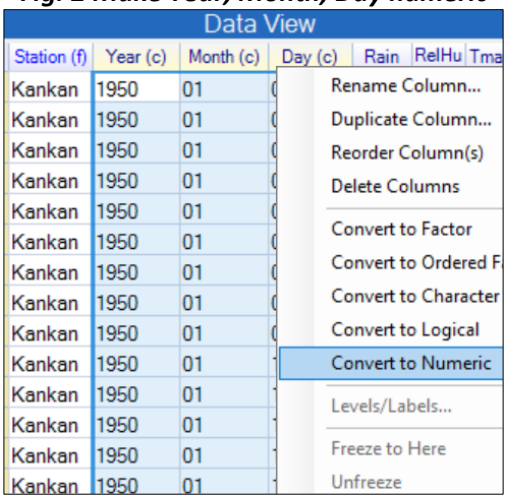


Fig. 2 Make Year, Month, Day numeric



** ***Right-click*** on the ***Station*** name and choose ***Convert to Factor,*** Fig. 1.

** ***Right-click*** also on the ***Year, Month and Day*** columns (you can mark them all at once) and ***convert them to numeric,*** Fig. 2***.***

Now we check the data are roughly as they should be. Surprises are not wanted!

Fig. 3 Choose the Summarise dialogue

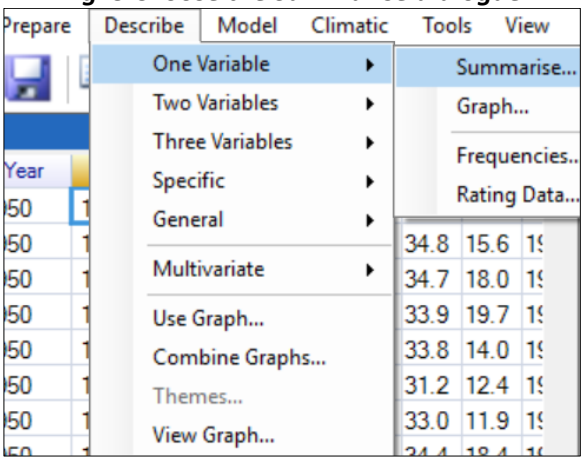
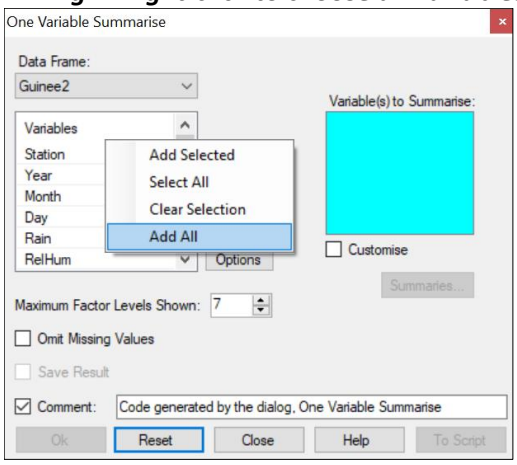


Fig. 4 Right-click to choose all variables



** Choose ***Describe > One Variable > Summarise***, Fig. 3.

Notice, Ok is not enabled in the dialogue. It first needs some variables to summarise.

** ***Right-click*** in the data selector, Fig. 4, and choose the option to ***Add All***. (Or just select all the variables and press the Add button.)

Fig. 5 Summarise dialogue completed

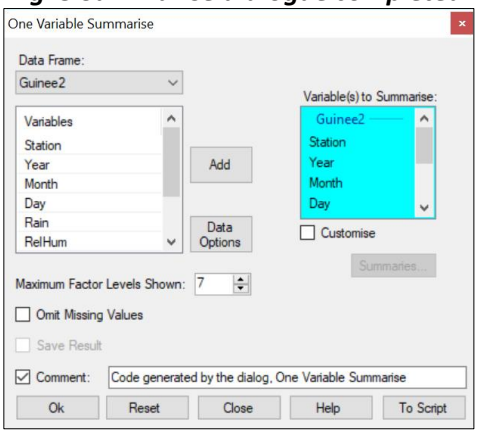
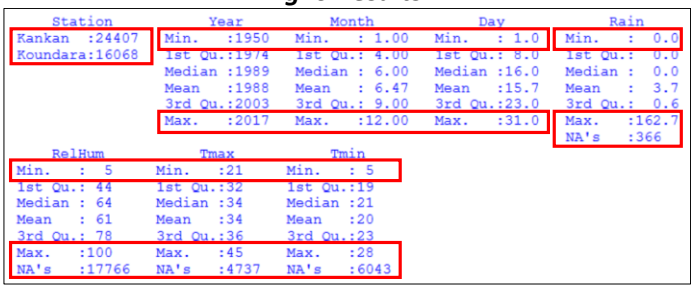


Fig. 6 Results



The dialogue is now completed and hence the Ok button is enabled.

** Press Ok.

** Examine the results – some interesting points are marked in red in Fig. 6. They include:

- The *Station* factor has just 2 levels. This is as expected, because we have 2 stations. There are more data for Kankan than Koundara and there are no missing values in this variable.
- For the *Rain* column, the minimum is zero (dry day) and the maximum is 163mm. These are plausible values. There are less than 400 missing values. They are denoted by NA in R.
- The *Year, Month, Day* columns are also as expected, for example Day is between 1 and 31 and there are no missing values.
- There are more missing values in the other 3 climatic elements. For these elements the minimum and maximum values are reasonable.
- There are *no really odd values*, like -99 that should have been made into missing values.

That's all comforting.

The next step is to make a single date variable.

** Choose ***Climatic > Dates > Make Date***, Fig. 7.

Fig. 7 Add a date variable

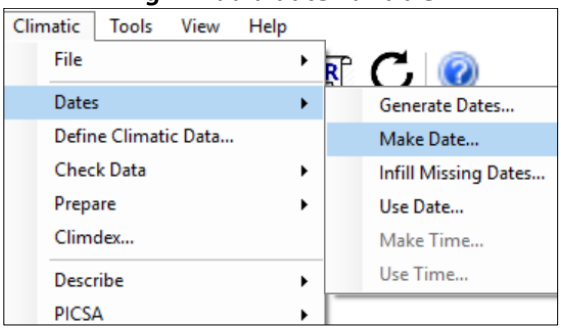
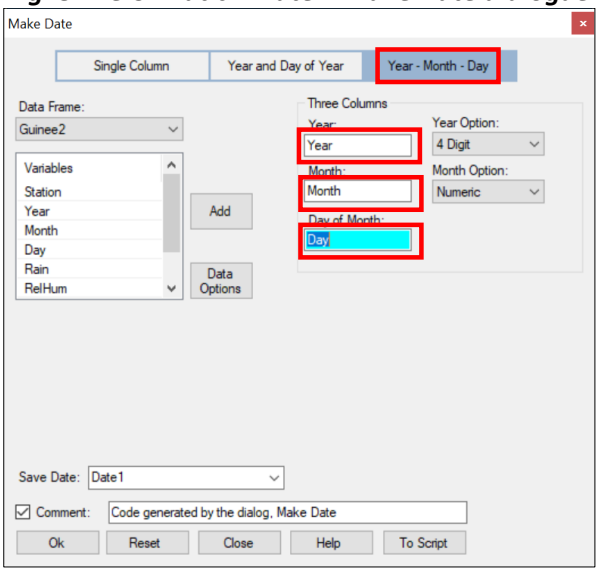


Fig. 8 The Climatic > Date > Make Date dialogue



** In the dialogue, Fig. 8 choose the ***Year – Month -Day*** button, because these 3 columns are in the current data frame.

** Complete the dialogue by ***adding*** the ***three columns***, as shown in Fig. 8. ***Press Ok.***

This has added a date column – of type (D) into the data frame.

The next step is to check whether any dates are missing from the file. This is not quite the same as missing values, but is when dates themselves are missing, perhaps whole years have been omitted from the file?

** Check the length of the data frame – currently 40475 rows (days) of data.

** Choose ***Climatic > Dates > Infill Missing Dates*** Fig. 9.

Fig. 9 Climatic >Dates > Infill

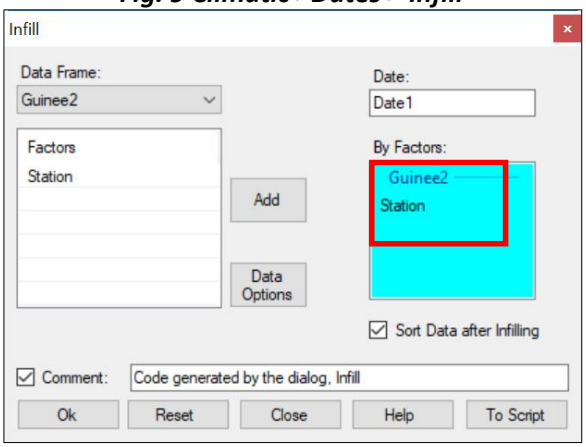
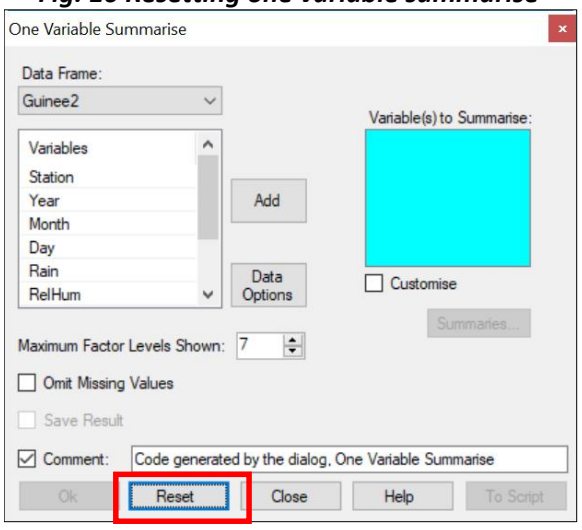


Fig. 10 Resetting one variable summarise



In Fig. 9 you should find the Date field was filled automatically.

** Click in the ***By Factors*** field and ***add the Station column***, Fig. 9.

** ***Click Ok***.

The length has now changed to 42063 rows. So about 1600 rows (days) have been added.

** Now use the ***Describe > One Variable > Summarise*** dialogue again. (Remember you also have a toolbar button to recall the last 10 dialogues.)

** Press the ***Reset*** button, Fig. 10.

** ***Right-click in the data selector*** (as you did before, Fig. 4) and ***Add All***.

** Press ***Ok***.

Fig. 11 Results from Describe > One Variable > Summarise

```
    Station          Year          Month            Day             Rain
Kankan  :24653   Min.   :1950   Min.   : 1.0   Min.   : 1.0   Min.   :  0.0
Koundara:17410   1st Qu.:1974   1st Qu.: 4.0   1st Qu.: 8.0   1st Qu.:  0.0
                 Median :1989   Median : 6.0   Median :16.0   Median :  0.0
                 Mean   :1988   Mean   : 6.5   Mean   :15.7   Mean   :  3.7
                 3rd Qu.:2003   3rd Qu.: 9.0   3rd Qu.:23.0   3rd Qu.:  0.6
                 Max.   :2017   Max.   :12.0   Max.   :31.0   Max.   :162.7
                 NA's   :1588   NA's   :1588   NA's   :1588   NA's   :1954
    RelHum           Tmax           Tmin           Date1
Min.   :  5    Min.   :21     Min.   : 5     Min.   :1950-01-01
1st Qu.: 44    1st Qu.:32     1st Qu.:19     1st Qu.:1974-05-25
Median : 64    Median :34     Median :21     Median :1988-10-16
Mean   : 61    Mean   :34     Mean   :20     Mean   :1987-12-03
3rd Qu.: 78    3rd Qu.:36     3rd Qu.:23     3rd Qu.:2003-03-09
Max.   :100    Max.   :45     Max.   :28     Max.   :2017-08-31
NA's   :19354  NA's   :6325   NA's   :7631
```

The results are shown in Fig. 11. Because the data have been infilled, there are now missing values in the Year, Month and Day columns. Something must be done about this. Fortunately, there are no missing values in the Date and the Station columns.

** Choose ***Climatic > Dates > Use Date***, see Fig. 4 for the menu.

** ***Complete*** it as shown in Fig. 12.

** Press ***Ok***.

This has generated 4 new columns, Fig. 13, for the year, the month (with labels), the day in the month, and the day of the year. The first 3 can replace the original columns which now have missing values after the infilling.

Fig. 12 Climatic > Dates > Use Date

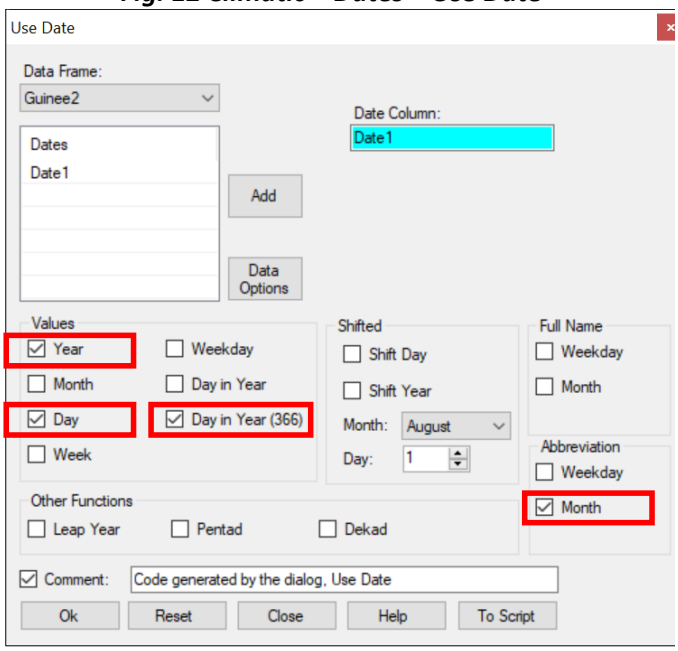
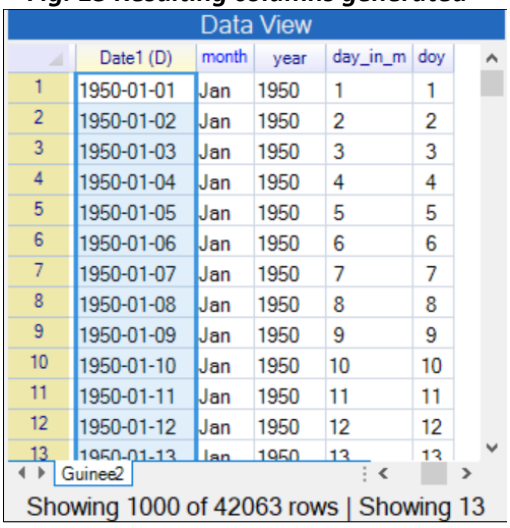


Fig. 13 Resulting columns generated



** ***Select the 3 original columns*** (now unwanted), ***right-click*** and choose ***Delete Columns***, Fig 14.

** Confirm the deletion, Fig. 15.

Fig. 14 Delete unwanted columns

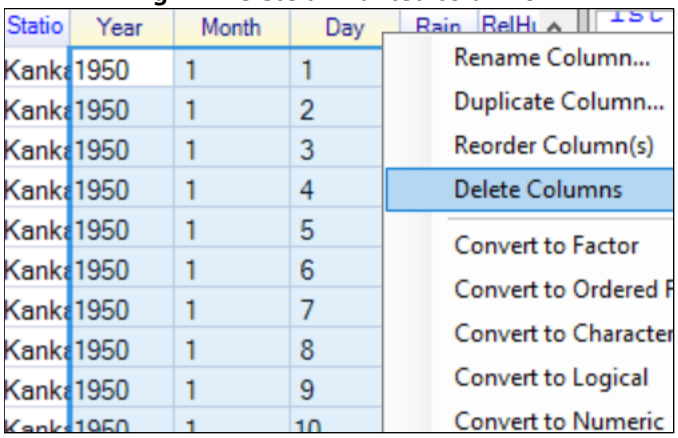
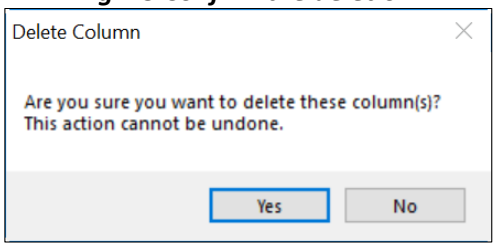


Fig. 15 Confirm the deletion



Finally, for this section, re-order the columns in the data frame, so the date columns are before the data.

** Right-click in the name field again, and choose ***Reorder Column(s)***, see Fig. 14.

** ***Reorder the columns*** as shown in Fig. 16.

** Press ***Ok***. The data frame is now as shown in Fig. 17.

Fig. 16 Reorder columns in the data frame

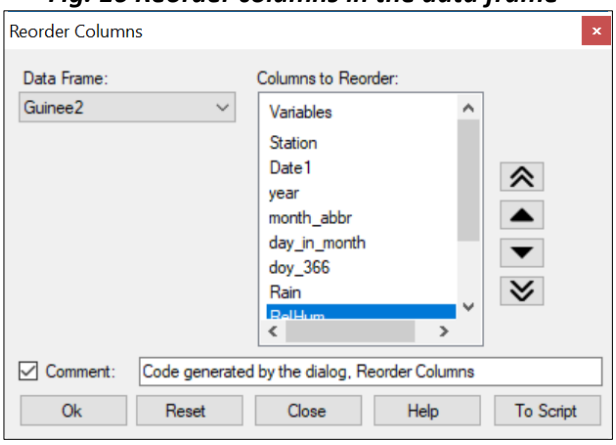
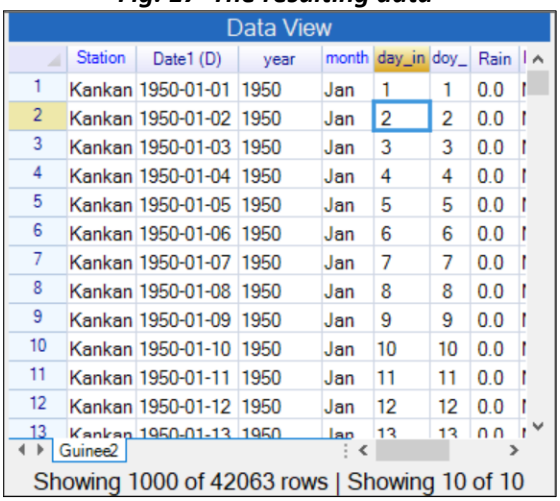


Fig. 17 The resulting data



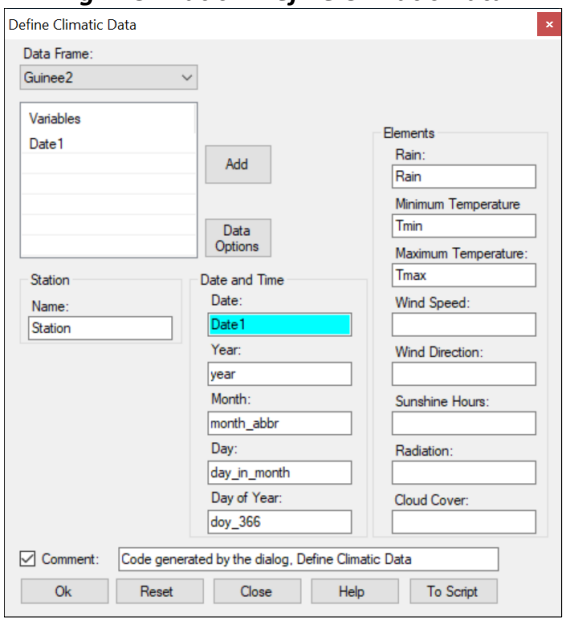
You could now save the data as described earlier, see section 3. But the next section is very short and could perhaps be done first.

5) Define the data as climatic

The final preparatory stage is to define the data as climatic.

** Use ***Climatic > Define Climatic Data***, Fig. 1.

Fig. 1 Climatic > Define Climatic Data



For this data set the dialogue was completed automatically. That is because R-Instat recognised the names of the columns. Otherwise the dialogue must be completed manually.

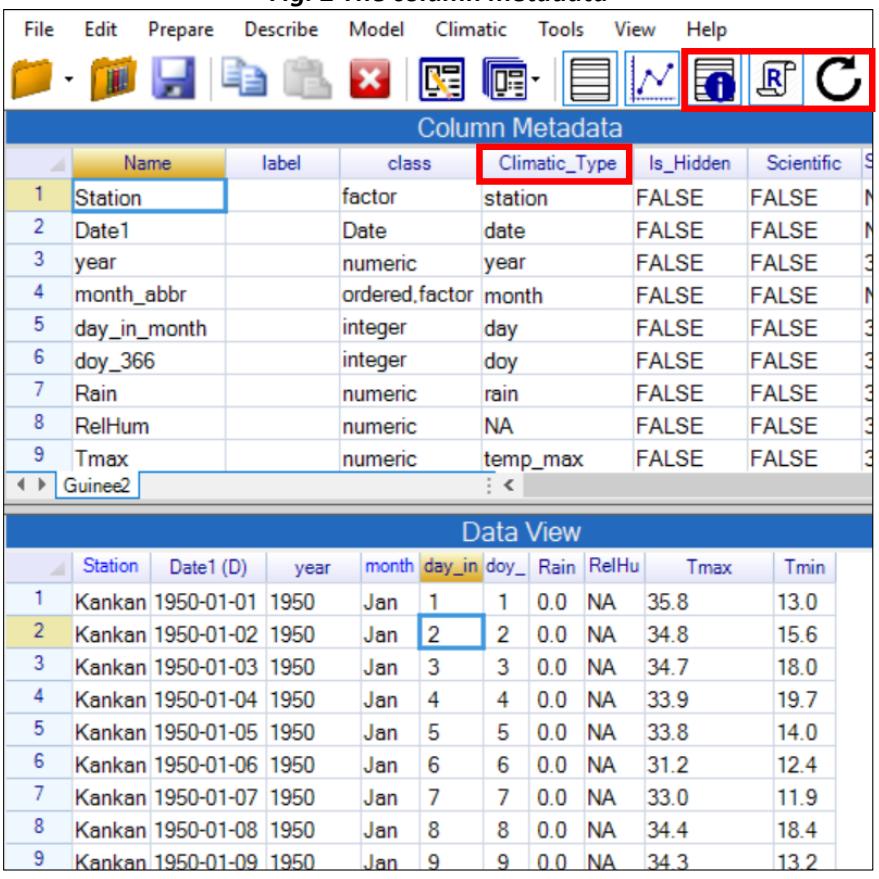
** Press ***Ok***.

It is unclear whether anything changed. So, we take this opportunity to introduce a third window in R-Instat.

So far you have seen a window for the data and another for the results.

** On the toolbar ***press the icon with an i*** – for information, Fig. 2.

Fig. 2 The column metadata



** ***Drag*** to make this metadata window bigger, as shown in Fig. 2

This window has a row that corresponds to each column in the Data window. What is new after the dialogue in Fig. 1 is that the column metadata includes the Climatic_Type information.

This simplifies the dialogues for climatic analyses in further sections of this guide.

Notice also that a label can be added to give further details about the contents of any column.

** Press the ***i button on the toolbar again*** to close the metadata window. Alternatively use the ***curly arrow*** to reset the windows back to their default positions.

Finally, in this section, save the data. This is now ready to start the analyses.

** Use ***File > Save As > Save Data As*** to give the dialogue shown in Fig. 3.

** Click on ***Browse***, Fig. 3 and choose where to save the data.

Fig. 3 File > Save As > Save Data As

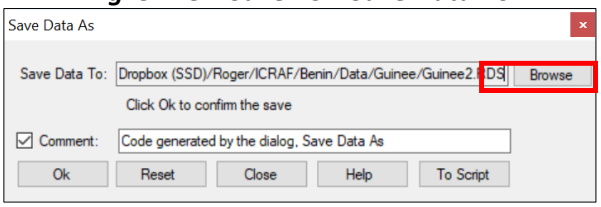
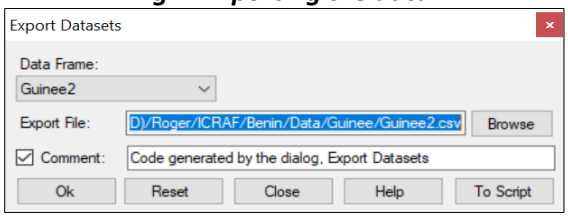


Fig. 4 Exporting the data



** Click ***Save*** on the resulting dialogue, to return to Fig. 3.

** Click ***Ok*** which is the step that actually saves the file.

** If you wish, you can also choose ***File > Export > Export Dataset***, Fig. 4. Then, click ***Browse***, ***Save***, and press ***Ok***.

The export has saved a csv file. This can be viewed in Excel, and later imported back into R-Instat. However, it does not save the metadata.

6) Checking data quality

** ***Continue with the data file*** from the sections above.

** (Otherwise use ***File > Open from Library > Load from Instat Collection > Browse > Climatic > Guinea*** and open the file ***Guinea2.RDS***)

Fig. 1 Climatic menu

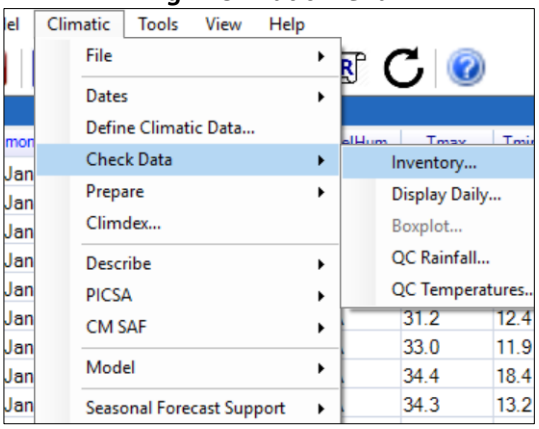
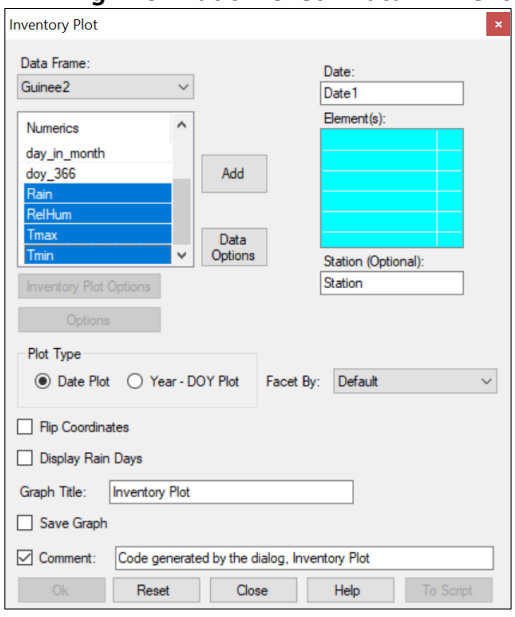


Fig. 2 Climatic > Check Data > Inventory



Section 4 showed menu options for the Dates and Section 5 used the Define Climatic Data dialogue. We now continue with dialogues that support checking the data.

** Choose ***Climatic > Check Data***, Fig. 1.

** Choose the ***Inventory*** dialogue from the menu in Fig. 1.

** In Fig. 2 Select the ***Elements receiver***, then ***choose the 4 climatic elements*** and press ***Add***.

** Press Ok.

Fig. 3 Inventory plot for Kankan and Koundara

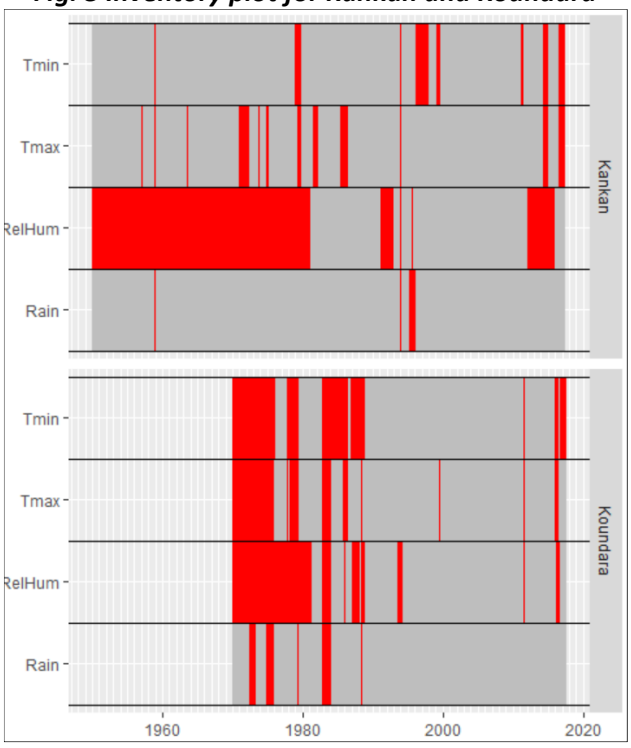
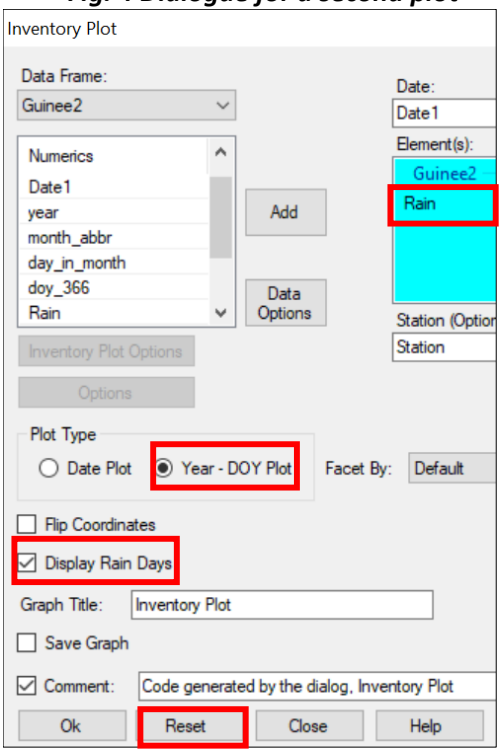


Fig. 4 Dialogue for a second plot



The top graph in Fig. 3 is for Kankan. The red indicates missing data and the record for the rainfall shows few missing values. The temperature data for Kankan also start in 1950. There are slightly more missing values, particularly for Tmax.

The rainfall data for Koundara start in about 1970, but the temperature data start a few years later. There are occasional missing periods in the early part of the rainfall record, but almost none later.

** Return to the ***Inventory dialogue***. (Use the toolbar icon.)

** Press the ***Reset*** button, Fig. 4.

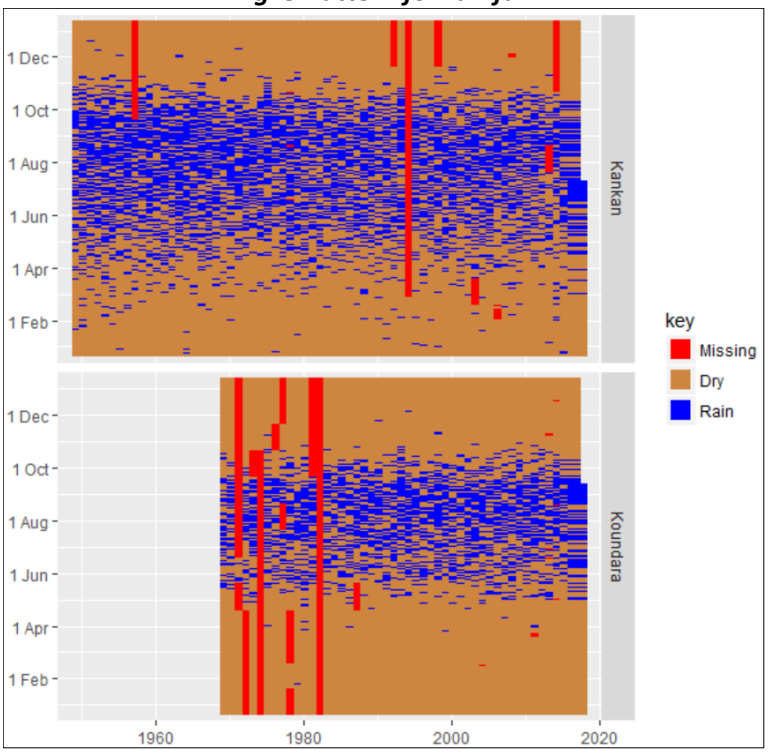
** Just put the ***Rain*** variable into the receiver, Fig. 4.

** Set the other 2 options as shown in Fig. 4 to ***display the rain days***, and to show ***days in the year***.

** Press Ok to give the results in Fig. 5.

The results in Fig. 5 indicate a single rainy season at both sites, with a longer season at Kankan. The data give an overall impression of good quality, e.g. there are no very odd years in the record.

Fig. 5 Pattern for rainfall



The next option in Climatic > Check data presents daily values in more detail.

This is slow for large amounts of data. We first *filter* to look just at the early years at Koundara.

Filtering is a powerful facility in R-Instat.

Put the cursor in ***the name field of the data* and ***right-click***, Fig. 6.

Fig. 6 Selecting Filter

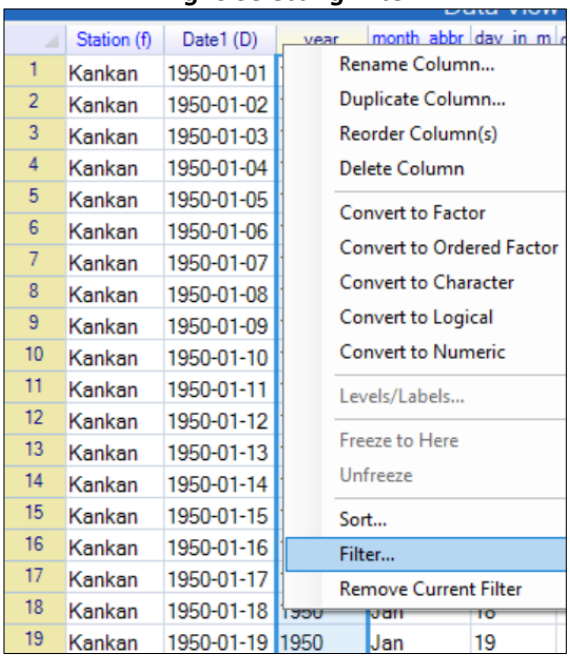
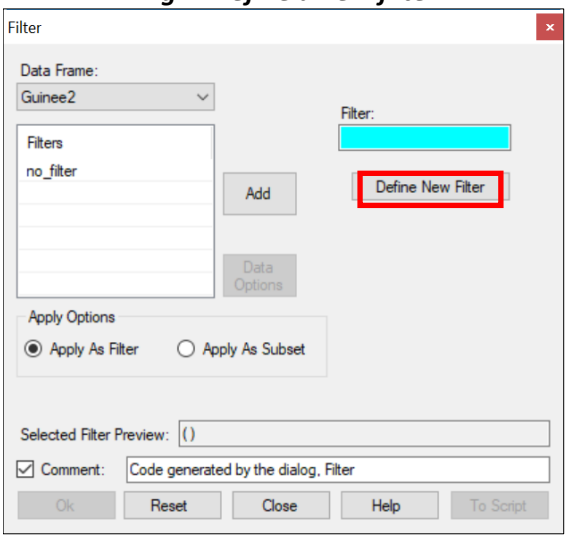


Fig. 7 Define a new filter



** In the Filter dialogue, ***click*** to define a ***new filter***, Fig. 7.

Fig. 8 Filter for level of a factor

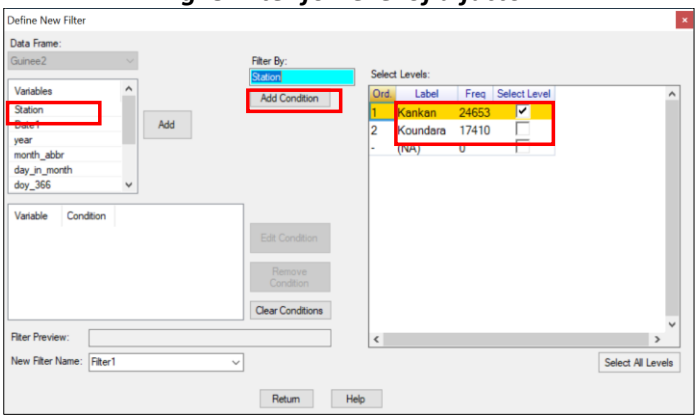
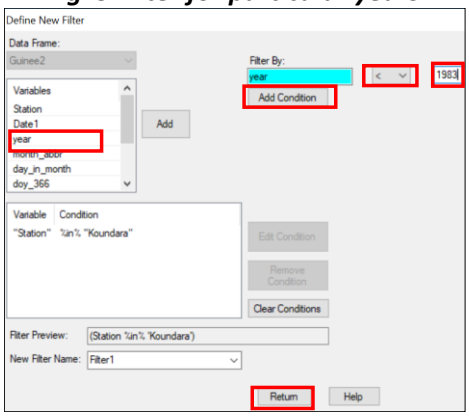


Fig. 9 Filter for particular years



First select the station.

** Choose to filter on the ***Station*** variable, Fig. 8..

** ***Choose Koundara***, rather than Kankan, Fig. 8. Then ***click to Add Condition***.

In Fig. 9, notice that the Station condition has now been added -It says “Station %in% 'Koundara'”.

Now we add a further condition as well.

** Choose the ***year*** column, Fig. 9.

** Set the ***condition to <*** and type the year as ***1983***, Fig. 9.

** ***Click to Add Condition***

** Now that the 2 conditions are added, press ***Return***.

Fig. 10 The filter has been defined

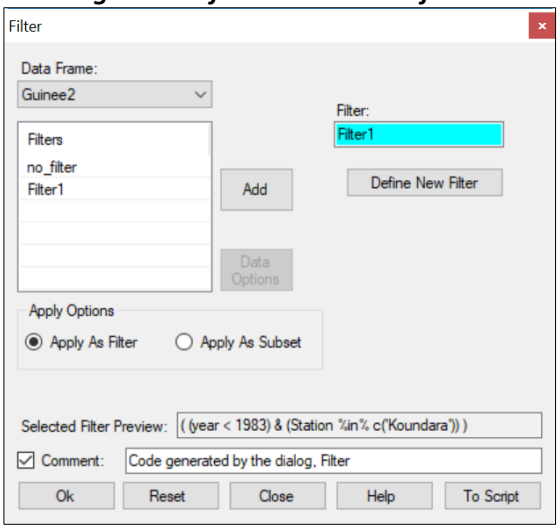
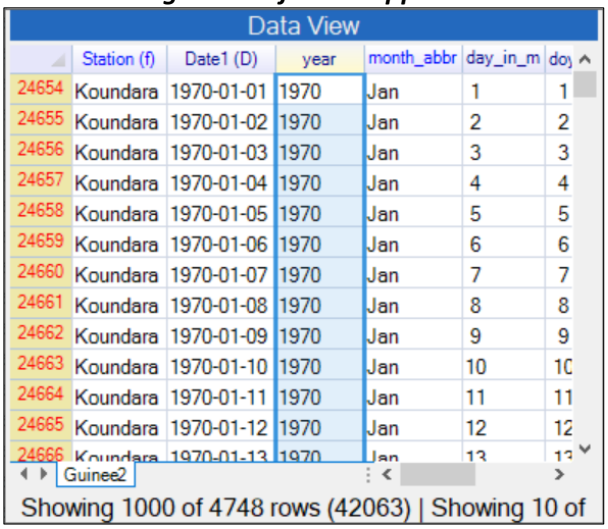


Fig. 11 The filter is applied



** Back on the main Filter dialogue, Fig. 10, ***press Ok***.

In Fig. 11 the first column is now in red, to indicate a filter is in operation. The data now starts with Koundara. Fig. 11 also indicates that just 4748 rows of data (out of the 42063 rows) have been selected for this stage in the analysis.

** Choose ***Climatic > Check Data***, as in Fig. 1 and choose the second option, ***Display Daily***.

** Choose to display the rainfall data, and ***complete the dialogue*** as shown in Fig. 12.

One year of the data is shown in Fig. 13, with the zero values and missing data shown clearly.

Fig. 12 The display daily data dialogue

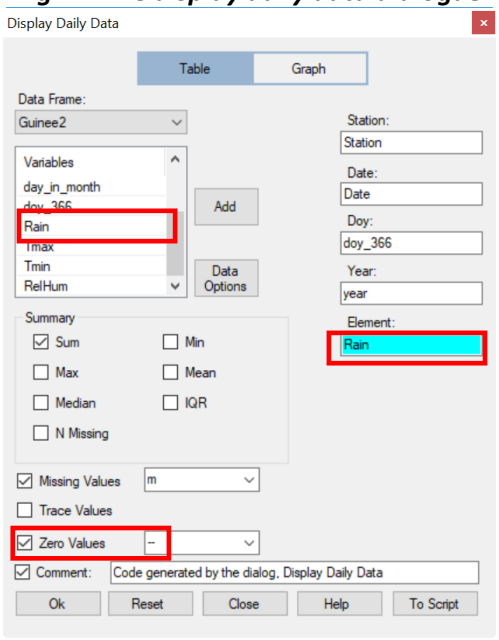
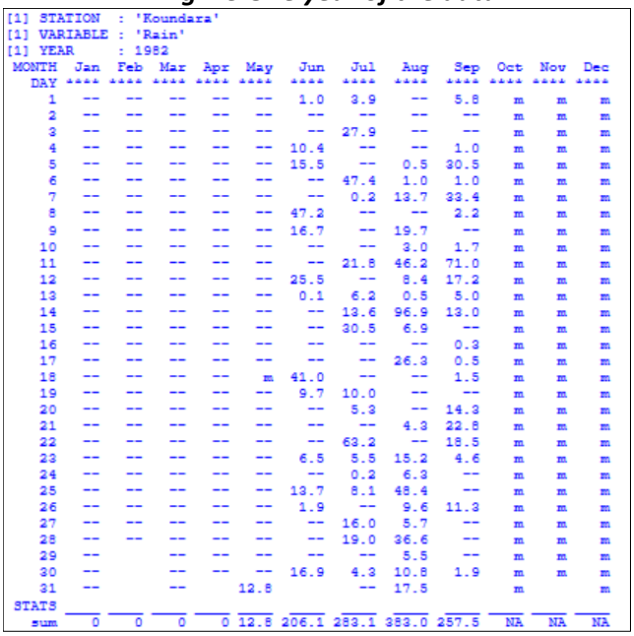


Fig. 13 One year of the data



```
[1] STATION  : 'Koundara'
[1] VARIABLE : 'Rain'
[1] YEAR     : 1982
 MONTH  Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
   DAY ****  ****  ****  ****  ****  ****  ****  ****  ****  ****  ****  ****
     1   --   --   --   --   --  1.0  3.9   --  5.8    m    m    m
     2   --   --   --   --   --   --   --   --   --    m    m    m
     3   --   --   --   --   --   --  27.9   --   --    m    m    m
     4   --   --   --   --   --  10.4   --   --  1.0    m    m    m
     5   --   --   --   --   --  15.5   --  0.5 30.5    m    m    m
     6   --   --   --   --   --   --  47.4  1.0  1.0    m    m    m
     7   --   --   --   --   --   --  0.2  13.7 33.4    m    m    m
     8   --   --   --   --   --  47.2   --   --  2.2    m    m    m
     9   --   --   --   --   --  16.7   --  19.7   --    m    m    m
    10   --   --   --   --   --   --   --  3.0  1.7    m    m    m
    11   --   --   --   --   --   --  21.8 46.2 71.0    m    m    m
    12   --   --   --   --   --  25.5   --  8.4 17.2    m    m    m
    13   --   --   --   --   --  0.1  6.2  0.5  5.0    m    m    m
    14   --   --   --   --   --   --  13.6 96.9 13.0    m    m    m
    15   --   --   --   --   --   --  30.5  6.9   --    m    m    m
    16   --   --   --   --   --   --   --   --  0.3    m    m    m
    17   --   --   --   --   --   --   --  26.3  0.5    m    m    m
    18   --   --   --   --   m  41.0   --   --  1.5    m    m    m
    19   --   --   --   --   --  9.7  10.0   --   --    m    m    m
    20   --   --   --   --   --   --  5.3   --  14.3    m    m    m
    21   --   --   --   --   --   --   --  4.3 22.8    m    m    m
    22   --   --   --   --   --   --  63.2   --  18.5    m    m    m
    23   --   --   --   --   --  6.5  5.5 15.2  4.6    m    m    m
    24   --   --   --   --   --   --  0.2  6.3   --    m    m    m
    25   --   --   --   --   --  13.7  8.1 48.4   --    m    m    m
    26   --   --   --   --   --  1.9   --  9.6 11.3    m    m    m
    27   --   --   --   --   --   --  16.0  5.7   --    m    m    m
    28   --   --   --   --   --   --  19.0 36.6   --    m    m    m
    29   --   --   --   --   --   --   --  5.5   --    m    m    m
    30   --   --   --   --   --  16.9  4.3 10.8  1.9    m    m    m
    31   --   --   --  12.8   --   --  17.5   --   --    m    m    m
 STATS
   sum    0    0    0    0 12.8 206.1 283.1 383.0 257.5   NA   NA   NA
```

We remove the filter before our next check on the data.

** Either ***go into the name field***, Fig. 6 and choose the last option ***Remove Current Filter***. Or go to the left-hand column (the one in red), Fig. 14, right-click and choose ***Remove Current Filter.***

Fig. 14 Right-click in the rows

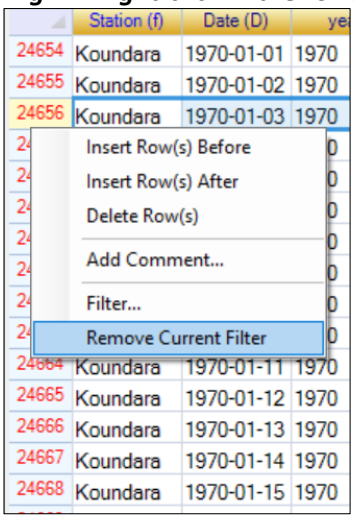
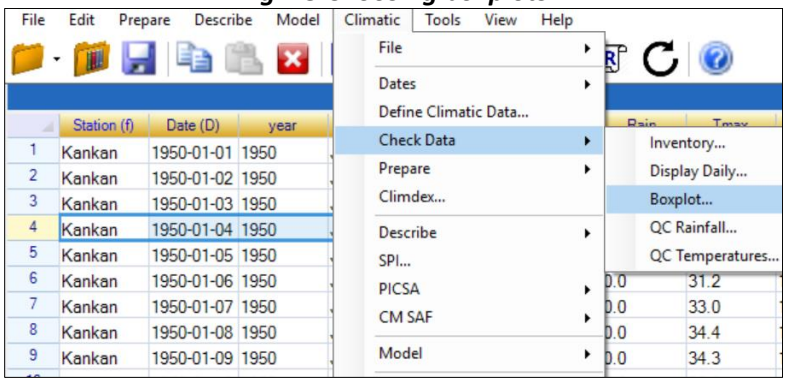


Fig. 15 Choosing boxplots



Boxplots are a useful way to explore the data and to check for data quality at the same time.

** Use Climatic > Check Data > Boxplots as shown in Fig. 15.

** Complete the dialogue ***as shown in Fig. 16***, so change the data to ***Tmax***, give 2 columns (facets) for the ***stations*** and show the seasonal pattern through plotting by ***Months***.

The result is shown in Fig. 17. The seasonal pattern is clear with the hottest months being March and April, and with Koundara being hotter then compared to Kankan.

There are also some surprising values to be investigated further and this is done later. A few are marked in red in Fig. 17.

Fig. 16 Boxplots by month for Tmax

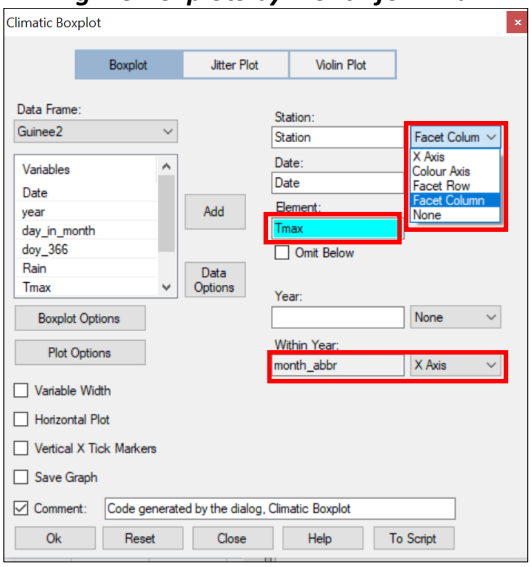
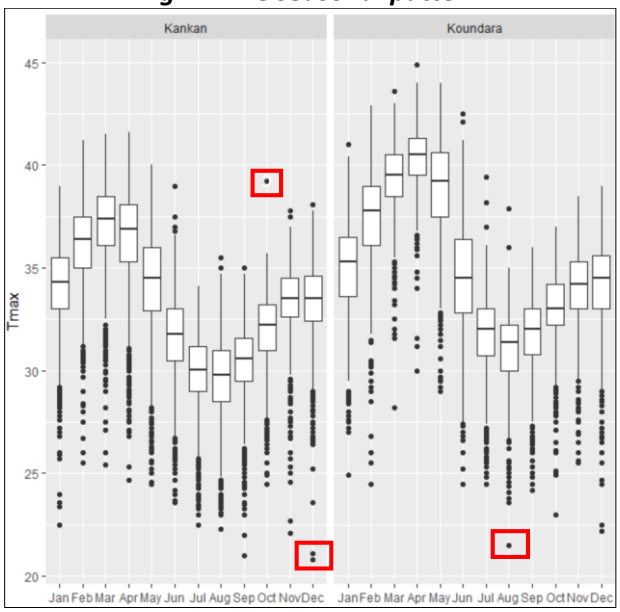


Fig. 17 The seasonal pattern



An alternative is to plot the data by year, i.e. in time-series order.

** Return to the last (climatic boxplot) dialogue.

** In Fig. 18 change the ***x-axis to the year column***. Make the boxplots of ***variable width*** and also make the ***x-axis labels vertical***. Change the ***Station*** from Facet Column, to ***Facet Row***.

Fig. 18 Boxplots by year for Tmax

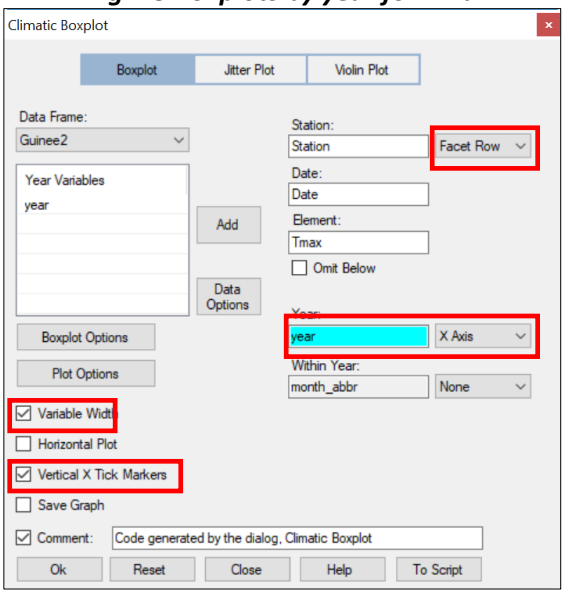
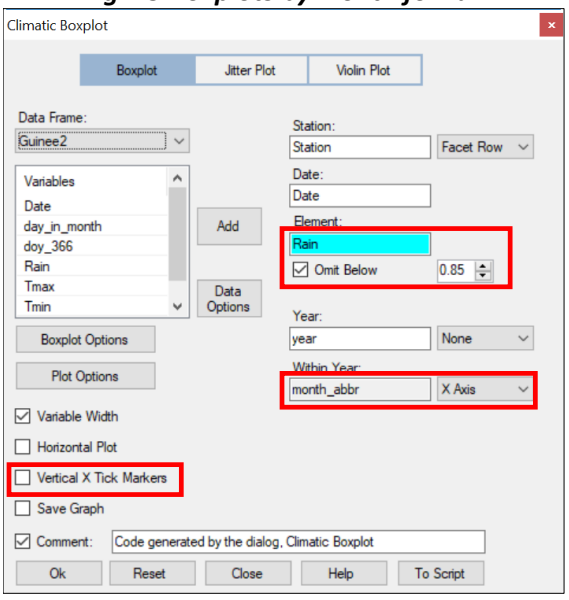


Fig. 19 Boxplots by month for rain



The result is shown below, Fig. 20.

Many other options are possible, but we illustrate with just one more

** Return to the last dialogue.

** In Fig. 19, change the variable to the ***rain column*** and tick the box to ***omit zero rainfalls***. Make the month into the x-axis again. The ***x-tick markers*** do not need to be vertical.

** Press Ok to give the results shown in Fig. 21. They show that the overall pattern of seasonality is the same. Kankan do have occasional rain in the December to March period, but this is very rare in Koundara.

Fig. 20 Boxplots for Tmax

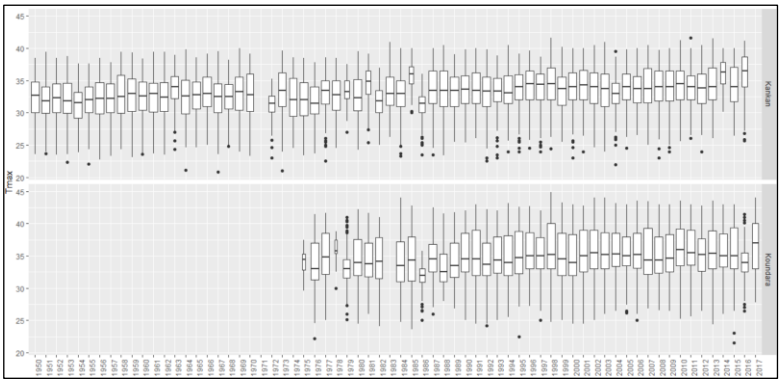
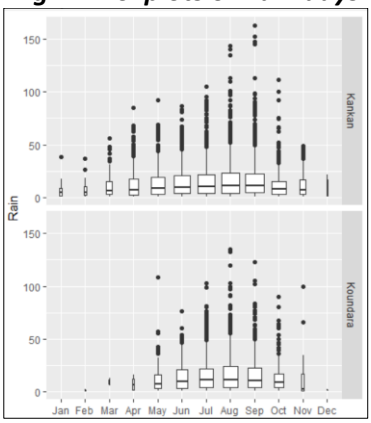


Fig. 21 Boxplots on rain days



Finally, in the Check data menu, we use the quality control dialogues. This is illustrated with checks for the temperature data.

** Use ***Climatic > Check Data > QC Temperatures***.

This dialogue is designed particularly for checks for Tmax and Tmin, though it can be used for any temperature data.

** Put Tamx and Tmin into the respective field, see Fig. 22.

** Tick the first 2 boxes and change the lower limit for Tmin (Element 2) to 5 degrees.

** The third option checks if 4 or more successive values of either Tmax or Tmin are the same value. Check this option also, see Fig. 22, leaving the option at the default value of 4 days

Fig. 22 Initial checks

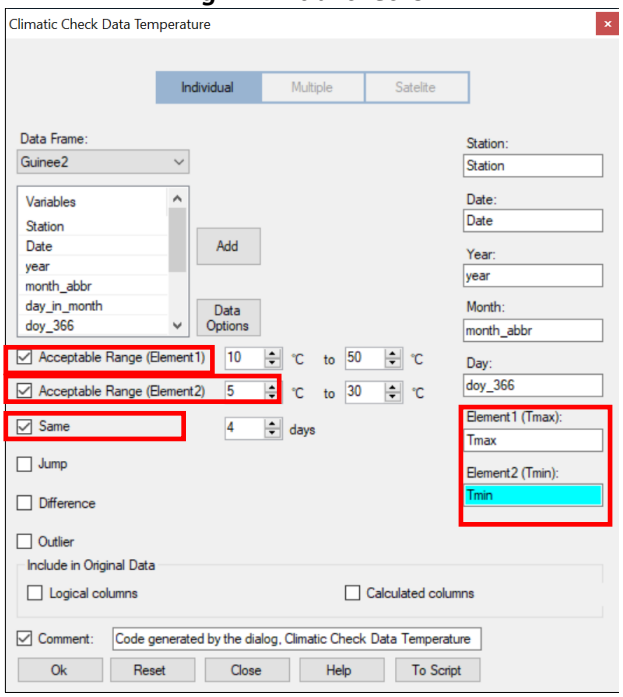
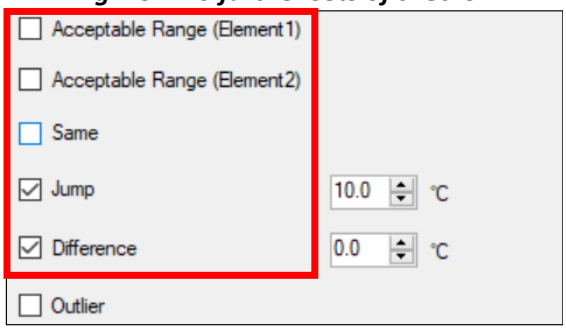
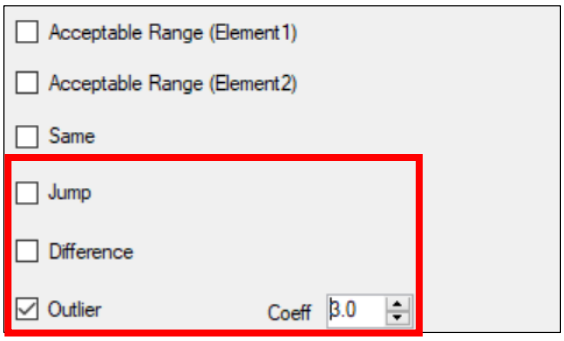


Fig. 23 Two further sets of checks





When you press Ok the results are put into another data frame, which is called QcTemp, see Fig. 24, below.

Before explaining the results, we conduct further checks.

** Return to the last dialogue.

** ***Uncheck the three options*** from Fig. 22 and ***check the next 2***, called ***Jump*** and ***Difference***.

The jump option shows days when the day-to-day difference for Tmax, or Tmin is large. The default is 10 degrees.

The difference option compares Tmax with Tmin and the default is to show all days when the difference is less than or equal to zero.

** ***Press Ok*** and you see that the results go into a further data frame, called QcTemp1.

** ***Return*** to the same dialogue again.

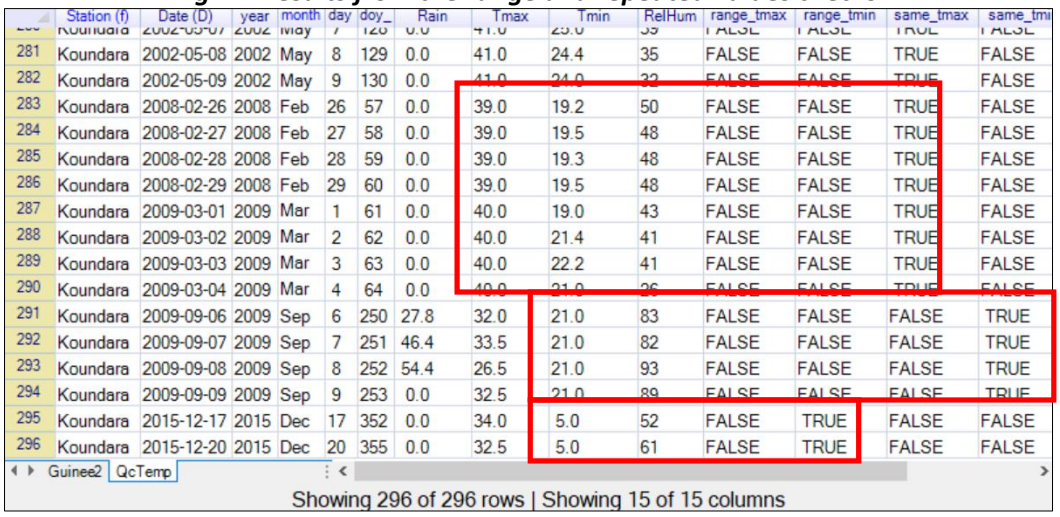
** ***Untick Jump and Difference*** and ***check the box for Outlier***.

** Change the Outlier value from 1.5 (the default for a boxplot) to 3.

** Press Ok to produce the filtered values in a new data frame called QcTemp2.

Now examine the resulting data frames in turn.

Fig. 24 Results from the range and repeated values checks.



	Station (f)	Date (D)	year	month	day	doy_	Rain	Tmax	Tmin	RelHum	range_tmax	range_tmin	same_tmax	same_tmi
281	Koundara	2002-05-08	2002	May	8	129	0.0	41.0	24.4	35	FALSE	FALSE	TRUE	FALSE
282	Koundara	2002-05-09	2002	May	9	130	0.0	41.0	24.0	33	FALSE	FALSE	TRUE	FALSE
283	Koundara	2008-02-26	2008	Feb	26	57	0.0	39.0	19.2	50	FALSE	FALSE	TRUE	FALSE
284	Koundara	2008-02-27	2008	Feb	27	58	0.0	39.0	19.5	48	FALSE	FALSE	TRUE	FALSE
285	Koundara	2008-02-28	2008	Feb	28	59	0.0	39.0	19.3	48	FALSE	FALSE	TRUE	FALSE
286	Koundara	2008-02-29	2008	Feb	29	60	0.0	39.0	19.5	48	FALSE	FALSE	TRUE	FALSE
287	Koundara	2009-03-01	2009	Mar	1	61	0.0	40.0	19.0	43	FALSE	FALSE	TRUE	FALSE
288	Koundara	2009-03-02	2009	Mar	2	62	0.0	40.0	21.4	41	FALSE	FALSE	TRUE	FALSE
289	Koundara	2009-03-03	2009	Mar	3	63	0.0	40.0	22.2	41	FALSE	FALSE	TRUE	FALSE
290	Koundara	2009-03-04	2009	Mar	4	64	0.0	40.0	21.0	36	FALSE	FALSE	TRUE	FALSE
291	Koundara	2009-09-06	2009	Sep	6	250	27.8	32.0	21.0	83	FALSE	FALSE	FALSE	TRUE
292	Koundara	2009-09-07	2009	Sep	7	251	46.4	33.5	21.0	82	FALSE	FALSE	FALSE	TRUE
293	Koundara	2009-09-08	2009	Sep	8	252	54.4	26.5	21.0	93	FALSE	FALSE	FALSE	TRUE
294	Koundara	2009-09-09	2009	Sep	9	253	0.0	32.5	21.0	89	FALSE	FALSE	FALSE	TRUE
295	Koundara	2015-12-17	2015	Dec	17	352	0.0	34.0	5.0	52	FALSE	TRUE	FALSE	FALSE
296	Koundara	2015-12-20	2015	Dec	20	355	0.0	32.5	5.0	61	FALSE	TRUE	FALSE	FALSE

Guinee2 QcTemp

Showing 296 of 296 rows | Showing 15 of 15 columns

Part of the first data frame, called QcTemp, is given above, Fig. 24.

There are 296 values to examine. The last 2 rows, in the red box, at the bottom of Fig. 24are December 17th and 20th 2015 and are the only occasions when the temperatures were out of the range, and they were minimum temperatures of just 5 degrees. The 4 rows above are September 6 to 9 2009 with a minimum temperature always 21.0 degrees.

The 8 rows above are February 26 to 29 2008, with a Tmax of 39.0 degrees, and then Match 1 to 4 has Tmax of 40.0 degrees.

One further feature of the "system" in R-Instat is that a logical column has been produced for each test, as shown in Fig. 24. This is TRUE for each day where the condition was not satisfied. This makes it easy to filter to investigate an individual condition in more detail. As an example we filter the data frame to look at the successive days when Tmin was the same.

We filtered before – see Fig. 6 and the following figures, so the ideas below should be familiar.

** Make sure you are in the ***data frame called QcTemp***.

** Then ***right click*** in the name row at the top and choose the option ***Filter***.

** In the subsequent dialogue choose the button ***Define new filter***.

** In the resulting sub-dialogue choose the column called ***same_tmin***.

** Make the condition that ***same_tmin == TRUE***.

** ***Rename*** the filter to be ***sameTmin***.

** Press ***Return*** to return to the main Filter dialogue, see Fig. 25.

** ***Press Ok*** and you see the first column in the data frame has turned red, to indicate a filter is in operation.

Fig. 25 A filter for Tmin

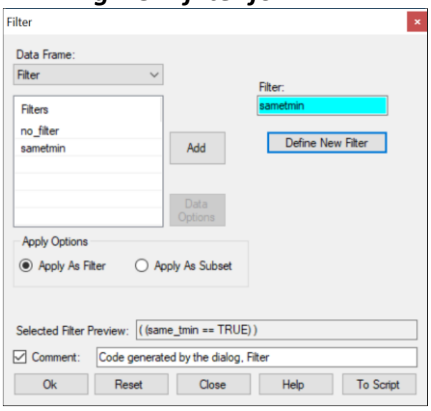
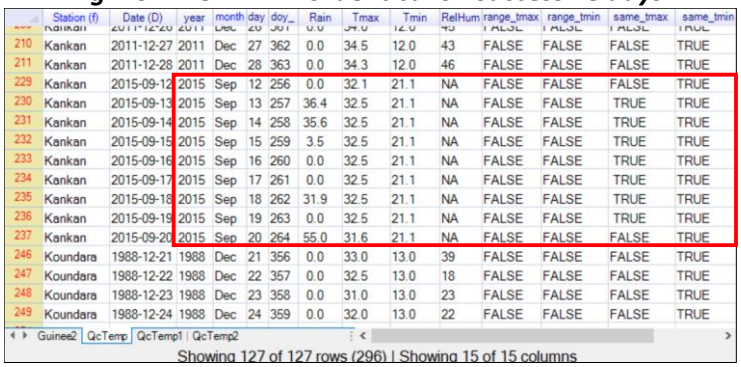


Fig. 26 When Tmin is identical on successive days



	Station (f)	Date (D)	year	month	day	doy_	Rain	Tmax	Tmin	RelHum	range_tmax	range_tmin	same_tmax	same_tmin
210	Kankan	2011-12-27	2011	Dec	27	362	0.0	34.5	12.0	43	FALSE	FALSE	FALSE	TRUE
211	Kankan	2011-12-28	2011	Dec	28	363	0.0	34.3	12.0	46	FALSE	FALSE	FALSE	TRUE
229	Kankan	2015-09-12	2015	Sep	12	256	0.0	32.1	21.1	NA	FALSE	FALSE	FALSE	TRUE
230	Kankan	2015-09-13	2015	Sep	13	257	36.4	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
231	Kankan	2015-09-14	2015	Sep	14	258	35.6	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
232	Kankan	2015-09-15	2015	Sep	15	259	3.5	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
233	Kankan	2015-09-16	2015	Sep	16	260	0.0	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
234	Kankan	2015-09-17	2015	Sep	17	261	0.0	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
235	Kankan	2015-09-18	2015	Sep	18	262	31.9	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
236	Kankan	2015-09-19	2015	Sep	19	263	0.0	32.5	21.1	NA	FALSE	FALSE	TRUE	TRUE
237	Kankan	2015-09-20	2015	Sep	20	264	55.0	31.6	21.1	NA	FALSE	FALSE	FALSE	TRUE
246	Koundara	1988-12-21	1988	Dec	21	356	0.0	33.0	13.0	39	FALSE	FALSE	FALSE	TRUE
247	Koundara	1988-12-22	1988	Dec	22	357	0.0	32.5	13.0	18	FALSE	FALSE	FALSE	TRUE
248	Koundara	1988-12-23	1988	Dec	23	358	0.0	31.0	13.0	23	FALSE	FALSE	FALSE	TRUE
249	Koundara	1988-12-24	1988	Dec	24	359	0.0	32.0	13.0	22	FALSE	FALSE	FALSE	TRUE

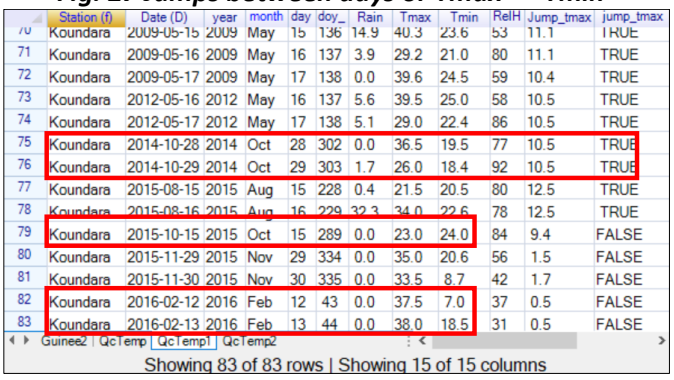
Guinee2 | QcTemp | QcTemp1 | QcTemp2

Showing 127 of 127 rows (296) | Showing 15 of 15 columns

From Fig. 26 we see there are now just 139 days to investigate. In Fig. 26 we have marked one interesting feature. In Kankan in September 2015 there are as many as 9 successive days when Tmin was recorded as 21.1 degrees. Also, on 7 of those days, the logical column for Tmax was also TRUE, i.e. Tmax was recorded as 32.5 degrees on each day.

We now look briefly at the results from the other checks, shown in Fig. 27 and Fig. 28.

Fig. 27 Jumps between days or Tmax <= Tmin

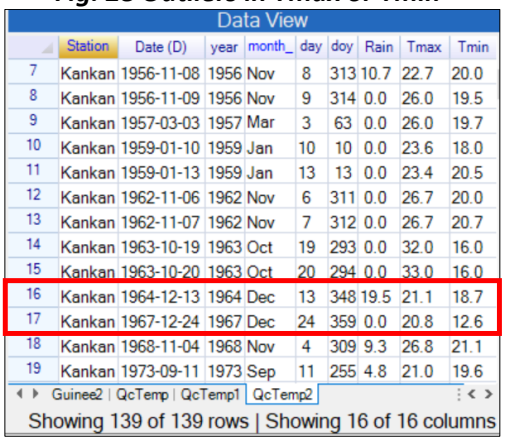


	Station (f)	Date (D)	year	month	day	doy_	Rain	Tmax	Tmin	RelH	Jump_tmax	jump_tmax
70	Koundara	2009-05-15	2009	May	15	136	14.9	40.3	23.6	53	11.1	TRUE
71	Koundara	2009-05-16	2009	May	16	137	3.9	29.2	21.0	80	11.1	TRUE
72	Koundara	2009-05-17	2009	May	17	138	0.0	39.6	24.5	59	10.4	TRUE
73	Koundara	2012-05-16	2012	May	16	137	5.6	39.5	25.0	58	10.5	TRUE
74	Koundara	2012-05-17	2012	May	17	138	5.1	29.0	22.4	86	10.5	TRUE
75	Koundara	2014-10-28	2014	Oct	28	302	0.0	36.5	19.5	77	10.5	TRUE
76	Koundara	2014-10-29	2014	Oct	29	303	1.7	26.0	18.4	92	10.5	TRUE
77	Koundara	2015-08-15	2015	Aug	15	228	0.4	21.5	20.5	80	12.5	TRUE
78	Koundara	2015-08-16	2015	Aug	16	229	32.3	34.0	22.6	78	12.5	TRUE
79	Koundara	2015-10-15	2015	Oct	15	289	0.0	23.0	24.0	84	9.4	FALSE
80	Koundara	2015-11-29	2015	Nov	29	334	0.0	35.0	20.6	56	1.5	FALSE
81	Koundara	2015-11-30	2015	Nov	30	335	0.0	33.5	8.7	42	1.7	FALSE
82	Koundara	2016-02-12	2016	Feb	12	43	0.0	37.5	7.0	37	0.5	FALSE
83	Koundara	2016-02-13	2016	Feb	13	44	0.0	38.0	18.5	31	0.5	FALSE

Guinee2 | QcTemp | QcTemp1 | QcTemp2

Showing 83 of 83 rows | Showing 15 of 15 columns

Fig. 28 Outliers in Tmax or Tmin



Data View

	Station	Date (D)	year	month_	day	doy	Rain	Tmax	Tmin
7	Kankan	1956-11-08	1956	Nov	8	313	10.7	22.7	20.0
8	Kankan	1956-11-09	1956	Nov	9	314	0.0	26.0	19.5
9	Kankan	1957-03-03	1957	Mar	3	63	0.0	26.0	19.7
10	Kankan	1959-01-10	1959	Jan	10	10	0.0	23.6	18.0
11	Kankan	1959-01-13	1959	Jan	13	13	0.0	23.4	20.5
12	Kankan	1962-11-06	1962	Nov	6	311	0.0	26.7	20.0
13	Kankan	1962-11-07	1962	Nov	7	312	0.0	26.7	20.7
14	Kankan	1963-10-19	1963	Oct	19	293	0.0	32.0	16.0
15	Kankan	1963-10-20	1963	Oct	20	294	0.0	33.0	16.0
16	Kankan	1964-12-13	1964	Dec	13	348	19.5	21.1	18.7
17	Kankan	1967-12-24	1967	Dec	24	359	0.0	20.8	12.6
18	Kankan	1968-11-04	1968	Nov	4	309	9.3	26.8	21.1
19	Kankan	1973-09-11	1973	Sep	11	255	4.8	21.0	19.6

Guinee2 | QcTemp | QcTemp1 | QcTemp2

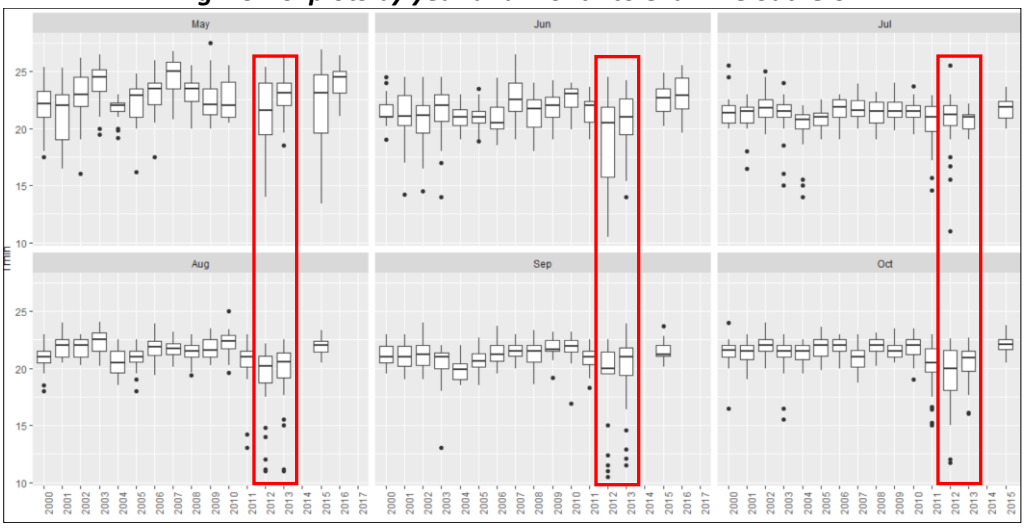
Showing 139 of 139 rows | Showing 16 of 16 columns

In Fig. 27 Tmax was 36.5 degrees C on October 28 2014 and 26.0 on the following day, a difference of 10.5 degrees. The second example indicated in Fig 27 was the only occasion in the record when Tmax was less than Tmin, which was on October 15th 2015. The third example indicated was a jump from 7 degrees to 18.5 degrees on February 12 and 13th 2016.

Fig 28 shows outliers using similar criteria to those in the boxplots, shown earlier in Fig 17 and 18. In Fig. 28 we have marked 2 days in December 1964 and 1967, where Tmax was low. These days were also marked as likely to need investigation in Fig. 17 earlier.

Finally, one surprise in the data frame of the outliers was that 39 of the 139 outliers were from Tmin in Kankan in the Summer months of 2012 and 2013. This was not obvious from boxplots of the type shown earlier. We leave as a challenge the detailed steps needed (filter plus boxplot) to produce the boxplot by year and month together, shown in Fig. 29.

Fig. 29 Boxplots by year and month to examine outliers



Appendix 1: Organising climatic data for R-Instat