



INTRODUCTORY TUTORIAL PART 2: A SECOND DATA SET

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

March 2021

CONTENTS

Contents	1
Chapter 1 — The Dodoma Data Set	2
1.1 Opening the New Data Set	2
1.2 Checking the Data Set	4
Chapter 2 — Preparing the Data	6
2.1 Applying a Filter	6
2.2 Producing Yearly Summaries	8
Chapter 3 — Analysing the Data	10
3.1 Producing graphs	10
3.2 Saving the Data	11
Chapter 4 Feedback and Reporting Bugs	12
References	13

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

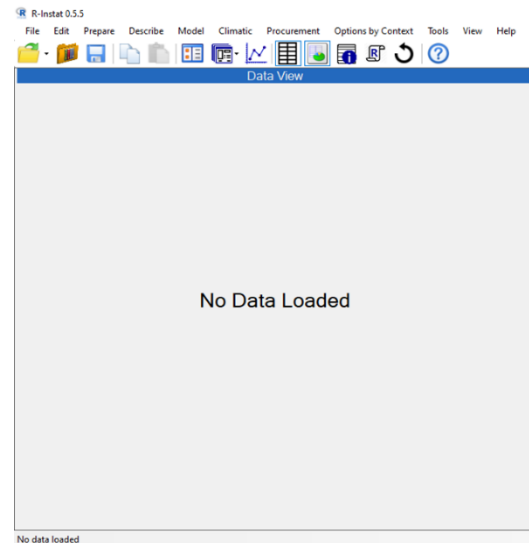
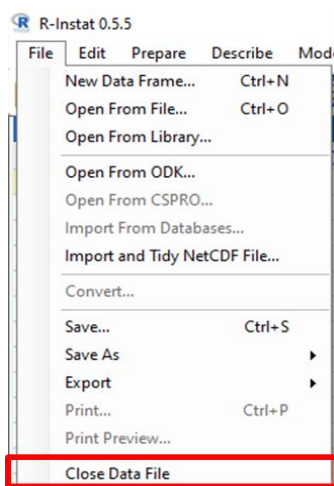
CHAPTER 1 — THE DODOMA DATA SET

This tutorial uses daily climatic data from Dodoma in Tanzania, from 1935 to 2013. 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. 15)
- You will be asked if you are sure. Respond **Yes**.

Fig. 15. Closing the previous data file

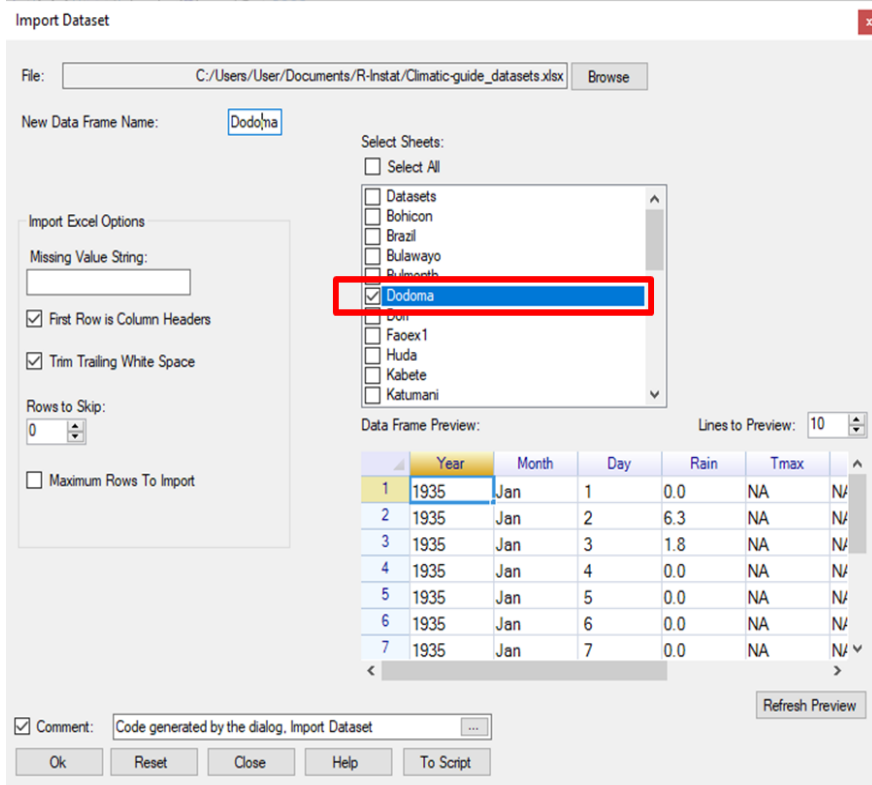
To start again



1.1 OPENING THE NEW DATA SET

- Use **File > Open from Library**. Take the option to **Load from Instat Collection** and then press **Browse**.
- Choose **Climatic** and select the Excel file **Climatic_guide_dataset**
- This Excel file has multiple sheets. Choose the one called **Dodoma**, see Fig. 16.

Fig 16: Opening the Dodoma sheet



The preview (Fig. 16) indicates this is ready to import. There are sensible names and the data starting on January 1st.

Fig 17: The Dodoma Daily Data

The screenshot shows the R-Instat 0.5.5 interface with the 'Data View' of the 'Dodoma' dataset. The table displays columns for Month (c), Day, Rain, Tmax, and Tmin, with data for January 1935. The data is as follows:

	Month (c)	Day	Rain	Tmax	Tmin
1	Jan	1	0.0	NA	NA
2	Jan	2	6.3	NA	NA
3	Jan	3	1.8	NA	NA
4	Jan	4	0.0	NA	NA
5	Jan	5	0.0	NA	NA
6	Jan	6	0.0	NA	NA
7	Jan	7	0.0	NA	NA
8	Jan	8	0.5	NA	NA
9	Jan	9	0.0	NA	NA
10	Jan	10	0.0	NA	NA
11	Jan	11	0.0	NA	NA
12	Jan	12	0.0	NA	NA
13	Jan	13	0.0	NA	NA
14	Jan	14	0.0	NA	NA
15	Jan	15	0.0	NA	NA
16	Jan	16	0.0	NA	NA

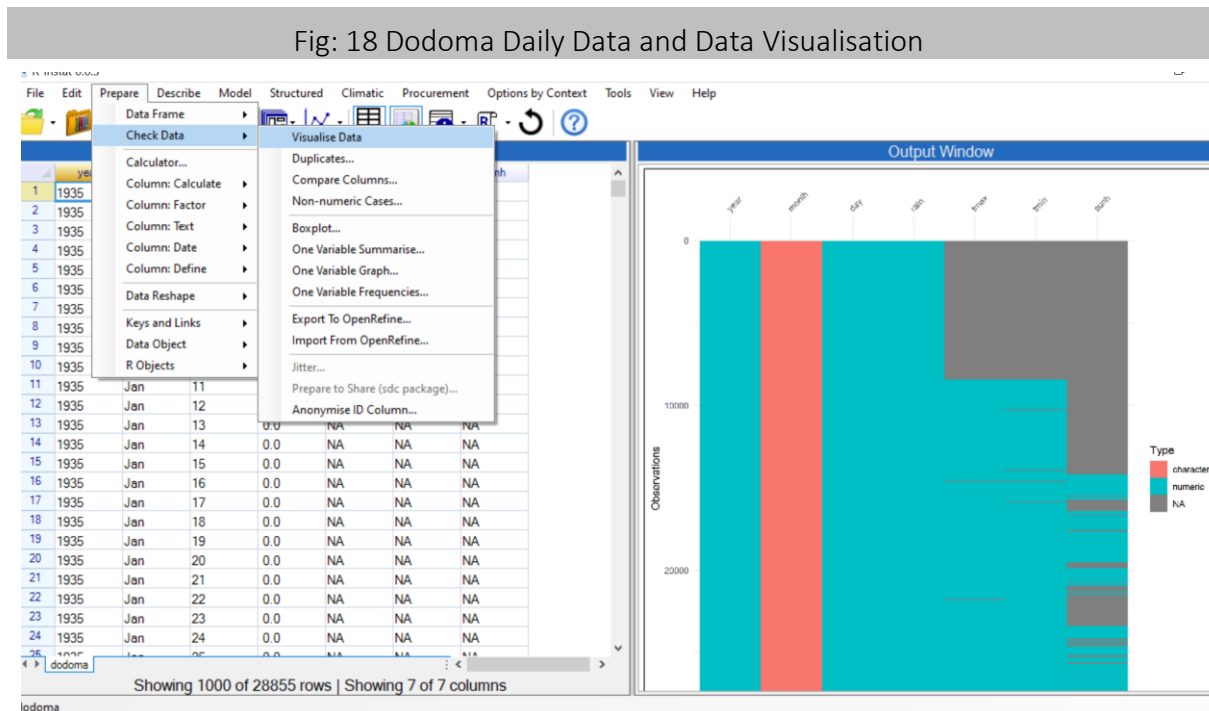
The data are shown in Fig. 17. These are daily data on rainfall, temperatures, and sunshine hours. There are 28,855 observations.

1.2 CHECKING THE DATA SET

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

- Now use Prepare > Check Data > Visualise Data.
- Just press Ok.

This gives a sort of “picture” of the data, see Fig. 18.



Grey is for the missing values. They indicate that the early years did not record the temperatures, and sunshine recording started later still. The figure also shows there aren't many missing values in the temperatures, once recording started, but the sunshine data is patchy.

- Return to the **Prepare > Check Data** menu.

The One Variable Summarise and Graph dialogues are repeated here. You used them from the Describe menu, but they are often also useful at the initial data checking stage.

- As before, with **One Variable > Summarise**, look at all the variables.

Fig: 19 Dodoma Data Summary

```
# Code generated by the dialog. One Variable Summarise
      year      month      day      rain
Min.   :1935   Length:28855   Min.   : 1.0   Min.   : 0.00
1st Qu.:1954   Class :character   1st Qu.: 8.0   1st Qu.: 0.00
Median :1974   Mode  :character   Median :16.0   Median : 0.00
Mean   :1974                                Mean   :15.7   Mean   : 1.57
3rd Qu.:1994                                3rd Qu.:23.0   3rd Qu.: 0.00
Max.   :2013                                Max.   :31.0   Max.   :119.80
                                         NA's   :91

      tmax      tmin      sunh
Min.   :15     Min.   : 8     Min.   : 0
1st Qu.:27     1st Qu.:15     1st Qu.: 8
Median :29     Median :17     Median :10
Mean   :29     Mean   :17     Mean   : 9
3rd Qu.:30     3rd Qu.:18     3rd Qu.:11
Max.   :36     Max.   :26     Max.   :14
NA's   :8631   NA's   :8703   NA's   :18451
```

This shows there were just **91 days when the rainfall** was missing. The visualisation (Fig 18) before indicated that there are a large number of missing values for the temperatures and sunshine. The summary confirms this, showing over **8 thousand values missing for temperature** and over **18 thousand missing from the sunshine data**.

On a positive note, there were over 20,000 days when the temperatures **were** measured, and more than 10,000 with sunshine data. Also notice there were no missing values on the year or day variables. That's comforting because, with a daily time series, you can't proceed if you don't know the date!

In hindsight, this presentation also implies that the diamonds data, in tutorial 1, did not have any missing values, or their presence would have been indicated.

The rainfall data in Fig. 17 are from 1935. The station added temperature records later.

- Use the **right-click** on the **bottom tab** and choose **View Data (the last option)** to view the whole data.
- Scroll down these data to confirm that the temperatures started from 1958.

This indicates that most of the 8 thousand missing temperature data in Fig. 18 are because of the later start of measuring these elements.

CHAPTER 2 — PREPARING THE DATA

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.

Our main work in this tutorial is to examine trends in the *annual* temperature data. This uses the Prepare menu because we don't yet have annual data. These are daily data.

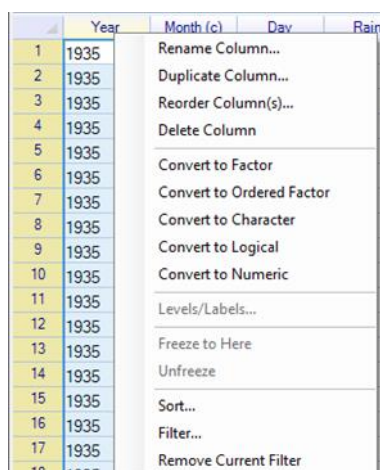
2.1 APPLYING A FILTER

First, filter the data, so it starts in 1958, when the temperatures start. The filter dialogue is available in Prepare > Data Frame > Filter

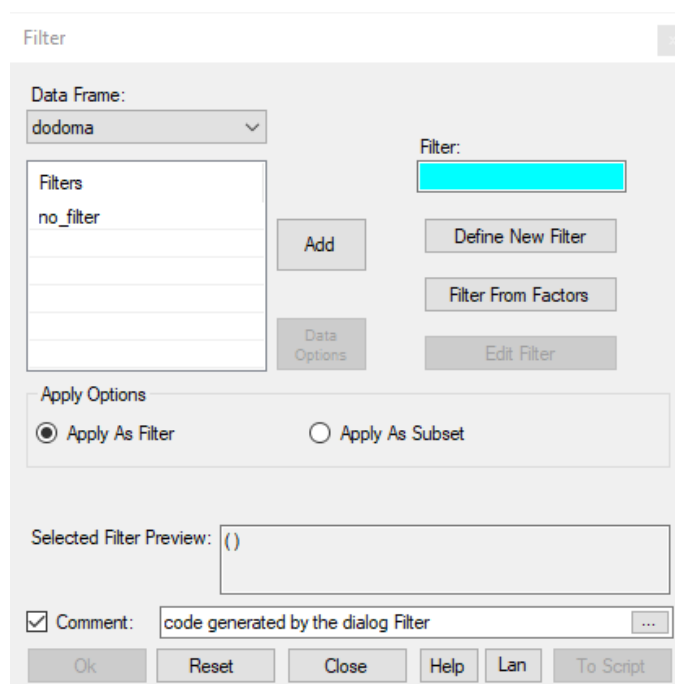
However, many common tasks from the **Prepare** menu are quickly accessible through a special **right-click menu** which is shown in Fig. 20.

- Put the cursor in the top row (with the names) and **right-click** (Fig. 20)
- Choose the **Filter dialogue** from this menu (Fig 20)

Fig20: The Right-Click Menu



Choose a Filter



- Click on the button in Fig. 20 to **Define New Filter**.
- In the sub-dialogue, choose the **year variable** and the condition that **is greater than 1957**. (Fig 21)
- **Add the condition** and optionally **give the condition a name**, we call it *from1958*. (Fig 21)

- Press **Return** and **OK**.

Fig 21: Define the Filter **And then apply it**

The image shows two side-by-side screenshots of R-Studio dialog boxes. The left screenshot is titled 'Define New Filter'. It shows a 'Data Frame' dropdown set to 'dodoma'. A list of variables is shown on the left, with 'year' selected. The 'Filter By:' field contains 'year' and a dropdown set to '1957'. There is a numeric keypad and an 'Add Condition' button. Below the keypad is a table with columns 'Variable' and 'Condition'. There are buttons for 'All Combined with &', 'Edit Condition', 'Remove Condition', and 'Clear Conditions'. A 'Filter Preview' field is empty, and a 'New Filter Name' dropdown is set to 'Filter1'. There are 'Return' and 'Help' buttons at the bottom. The right screenshot is titled 'Filter'. It shows the same 'Data Frame' dropdown. A list of filters is shown on the left, with 'Filter1' selected. There are buttons for 'Add', 'Define New Filter', 'Filter From Factors', and 'Edit Filter'. Below this is an 'Apply Options' section with two radio buttons: 'Apply As Filter' (selected) and 'Apply As Subset'. A 'Selected Filter Preview' field contains the expression '(year > 1957)'. There is a 'Comment' field with the text 'code generated by the dialog Filter'. At the bottom are buttons for 'Ok', 'Reset', 'Close', 'Help', 'Lan', and 'To Script'.

- 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.

On the left-hand-side the row numbers are now in red – which indicates a filter is in operation. The data now “start” in row 8402 with the temperature data., i.e. 1st January 1958.

Fig 22: Filtered Data

Data View								
	year	month (c)	day	rain	tmax	tmin	sunh	yr_temp (L)
8402	1958	Jan	1	0.0	28.6	18.7	NA	TRUE
8403	1958	Jan	2	0.0	29.7	18.8	NA	TRUE
8404	1958	Jan	3	0.0	29.7	17.6	NA	TRUE
8405	1958	Jan	4	7.1	30.5	18.8	NA	TRUE
8406	1958	Jan	5	8.9	31.2	19.2	NA	TRUE
8407	1958	Jan	6	2.0	31.1	19.1	NA	TRUE
8408	1958	Jan	7	0.0	27.2	18.1	NA	TRUE
8409	1958	Jan	8	0.0	28.9	18.8	NA	TRUE
8410	1958	Jan	9	0.0	30.0	16.7	NA	TRUE
8411	1958	Jan	10	0.0	30.1	17.3	NA	TRUE
8412	1958	Jan	11	0.0	31.2	19.3	NA	TRUE
8413	1958	Jan	12	0.0	31.2	19.1	NA	TRUE
8414	1958	Jan	13	0.0	32.1	18.3	NA	TRUE
8415	1958	Jan	14	0.0	31.8	18.6	NA	TRUE
8416	1958	Jan	15	0.0	32.9	18.3	NA	TRUE
8417	1958	Jan	16	0.0	33.6	17.8	NA	TRUE
8418	1958	Jan	17	0.0	34.1	19.2	NA	TRUE
8419	1958	Jan	18	0.3	32.6	18.9	NA	TRUE
8420	1958	Jan	19	0.0	33.3	19.4	NA	TRUE
8421	1958	Jan	20	0.0	32.7	20.0	NA	TRUE
8422	1958	Jan	21	0.0	33.2	19.9	NA	TRUE
8423	1958	Jan	22	0.4	31.8	18.8	NA	TRUE

- Return to the **Prepare > Check Data > One Variable Summarise** dialogue (You can use the toolbar icon for recalling the last ten dialogues).
- Press **Ok**.

Fig: 23 Summary of Filtered Data

```

      year      month      day      date
Min.   :1958    1       :1736  Min.   : 1.0   Min.   :1958-01-01
1st Qu.:1972    3       :1736  1st Qu.: 8.0   1st Qu.:1972-01-01
Median :1986    5       :1736  Median :16.0  Median :1985-12-31
Mean   :1986    7       :1736  Mean   :15.7  Mean   :1985-12-31
3rd Qu.:2000    8       :1736  3rd Qu.:23.0  3rd Qu.:1999-12-31
Max.   :2013   10      :1736  Max.   :31.0  Max.   :2013-12-31
      12      :1736
      4       :1680
      6       :1680
      (Other):4942

      month_abbr  doy_366  rain      tmax      tmin
Jan   :1736     Min.   : 1   Min.   : 0.00  Min.   :15.2  Min.   : 7.9
Mar   :1736     1st Qu.: 93  1st Qu.: 0.00  1st Qu.:27.4  1st Qu.:15.1
May   :1736     Median :184  Median : 0.00  Median :29.0  Median :17.2
Jul   :1736     Mean   :184  Mean   : 1.58  Mean   :28.9  Mean   :16.8
Aug   :1736     3rd Qu.:275  3rd Qu.: 0.00  3rd Qu.:30.5  3rd Qu.:18.5
Oct   :1736     Max.   :366  Max.   :119.80  Max.   :35.5  Max.   :25.5
Dec   :1736     NA's   :91   NA's   :230   NA's   :302
Apr   :1680
Jun   :1680
(Other):4942

```

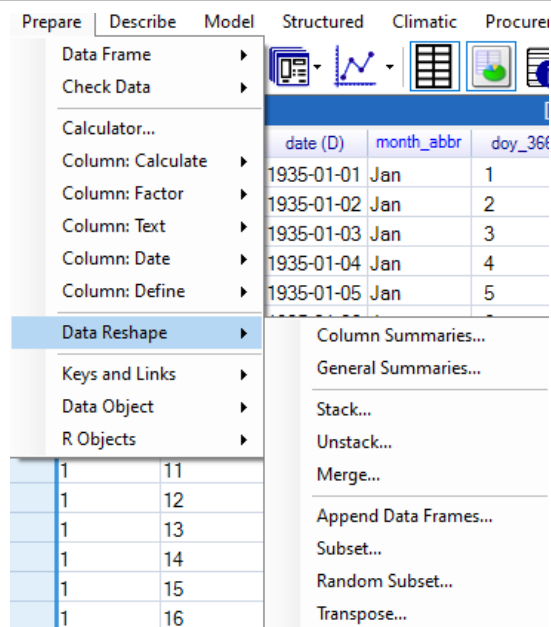
The results (Fig 23) include that there were only 230 missing days in tmax, from when temperatures were measured. A few more were missing in tmin.

2.2 PRODUCING YEARLY SUMMARIES

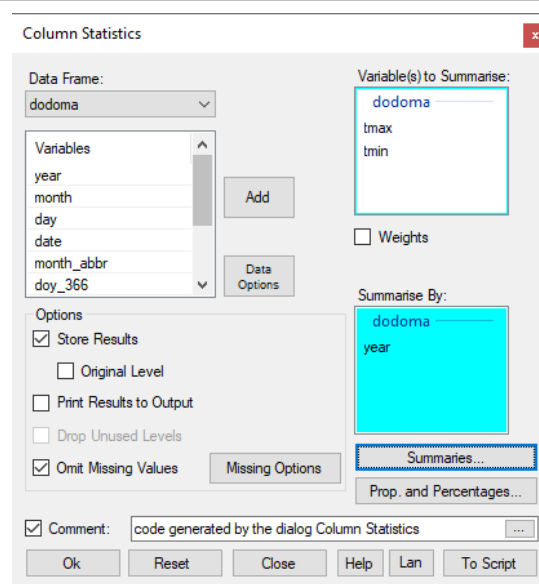
The daily data are now ready to be summarized to produce the yearly means.

- Open the **Prepare > Column: Reshape > Column Summaries** dialogue (Fig 24)

Fig 24: Menu for Column Summaries



With the resulting dialog



- Complete the dialogue (Fig. 24), 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. 25)

- Complete the sub-dialogue as shown in Fig 25, i.e. with only two summaries for the **N Not Missing** and the **Mean**. Then press **Return** and **OK** to produce the summaries.

Fig 25: Summaries Sub-dialog

With the Resulting Data

The screenshot shows the 'Summaries' dialog box on the left and the 'Data View' table on the right. The dialog box has several sections with checkboxes:

- Common:** N Non Missing, N Total, N Missing, Mode, n_distinct
- All but (unordered) Factor:** Minimum, Maximum, Range
- Numeric:** Sum, Mean, Median, Sd, Var
- Quartiles:** Lower Quartile, Upper Quartile

The 'Data View' table shows the following data:

	year (f)	mean_tmax	count_non_	mean_tmin	count_non_
1	1958	29.0	365	16.1	365
2	1959	28.7	365	16.3	365
3	1960	29.0	365	15.9	365
4	1961	29.3	365	17.1	365
5	1962	29.0	365	16.1	365
6	1963	28.5	363	16.0	331
7	1964	28.9	360	15.7	359
8	1965	28.8	363	16.0	354
9	1966	29.1	365	16.6	364
10	1967	28.5	365	16.7	365
11	1968	27.9	366	15.6	366
12	1969	29.7	365	17.0	365
13	1970	28.6	365	16.5	365
14	1971	28.5	365	16.3	365
15	1972	28.8	366	16.6	366
16	1973	29.5	362	16.6	334
17	1974	28.8	304	16.2	304
18	1975	29.1	365	16.8	365
19	1976	29.2	366	16.9	366
20	1977	28.7	364	17.1	364
21	1978	28.5	365	16.8	322
22	1979	28.4	365	16.5	364
23	1980	28.9	366	17.0	366
24	1981	29.1	365	16.7	365
25	1982	29.0	365	16.8	365

Fig. 25 The results are in a new sheet – or data frame in R terminology. There are just 56 rows, with one for each year. And they start in 1958. (Video: Indicate the number of years at the bottom of the data frame.)

So, now the daily data are in the first data frame and the annual summaries are in a second one.

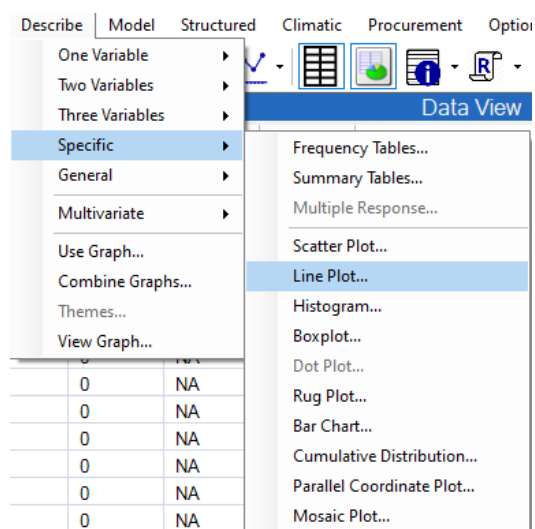
That was easy. The Prepare stage has not taken too long.

CHAPTER 3 — ANALYSING THE DATA

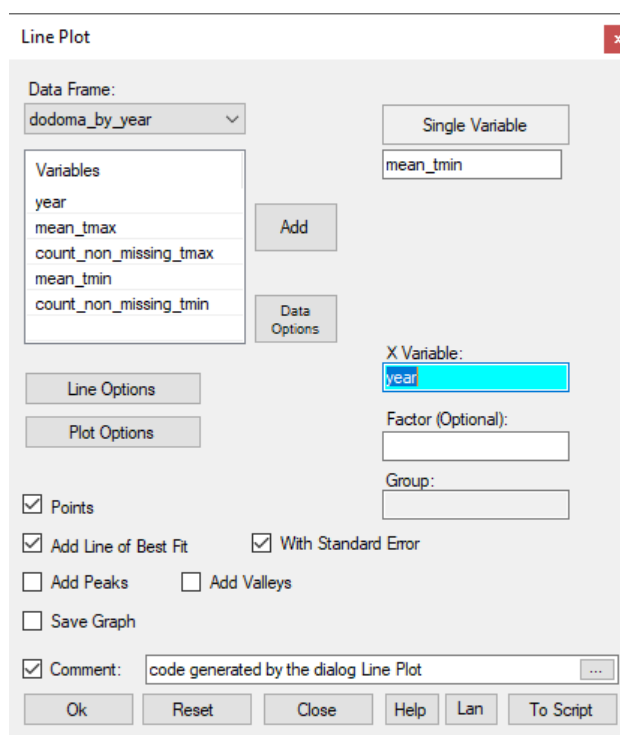
3.1 PRODUCING GRAPHS

- Use **Describe > Specific > Line Plot**, Fig. 26.
- Complete the dialogue as shown in Fig. 26 for the with **mean_tmin** as the y and **year** as the x.
- Add **the points**, and **the fitted line**.
- Press **OK**.

Fig 26: The line plot menu

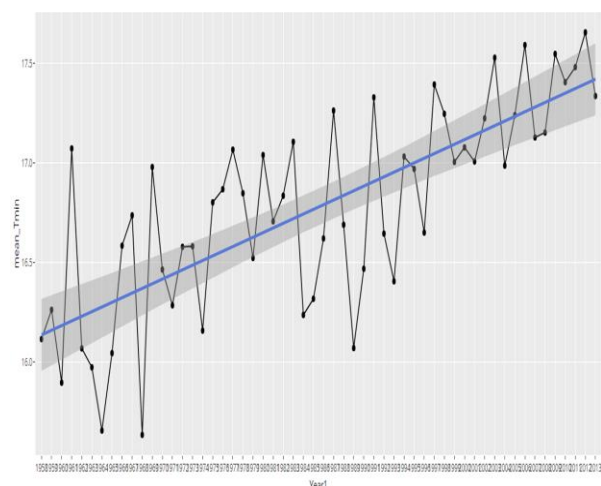


And the dialog

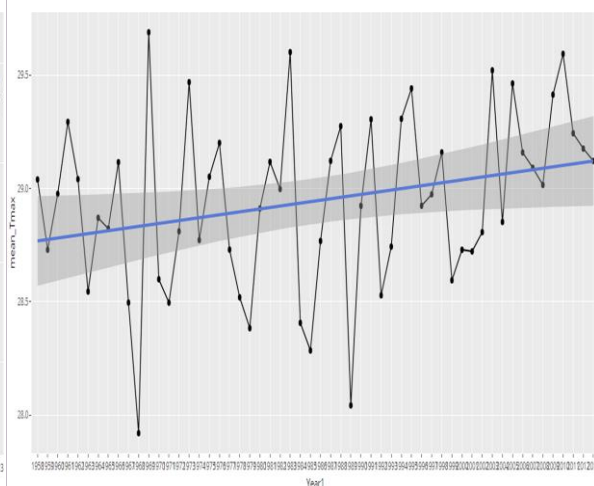


The resulting graph is shown in Fig. 26.

Fig 26: The Graph for t-min



And for t-max



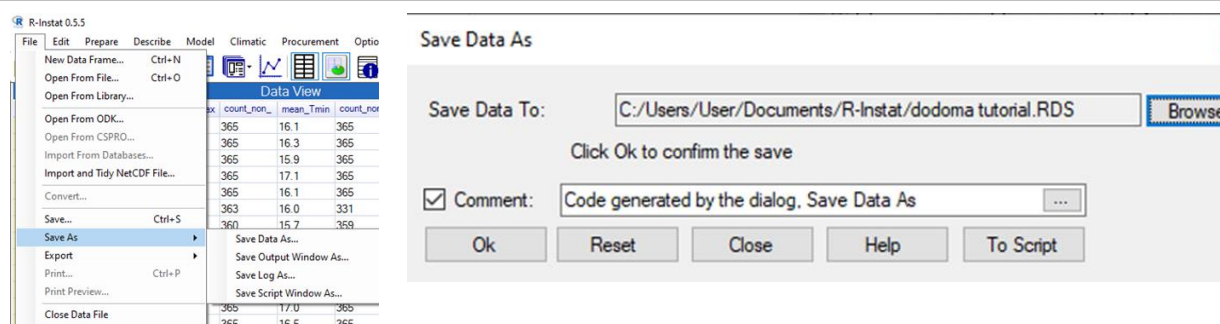
The result indicates that maximum temperatures have not increased so much.

3.2 SAVING THE DATA

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

- Use the **File > Save As** dialog, Fig. 27. Choose the option **Save Data As**.
- Press on **Browse** in the dialogue, Fig. 27. Choose a suitable directory and name. Press **OK** when you return to the Save Data dialogue.

Fig 27: Saving the Data set



The RDS extension is added, to signify it is saved as an R data file. This is a "silver lining", 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.

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.

CHAPTER 4 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.

REFERENCES

R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from <https://www.R-project.org/>.

Stern, R. D., Rijks, D., Dale, I. C., & Knock, J. (2006). *Instat Climatic Guide*.

Wikipedia contributors

Wikipedia contributors (2019)

Wikipedia contributors (2019). R (programming language), *Wikipedia, The Free Encyclopedia*. [https://en.wikipedia.org/w/index.php?title=R_\(programming_language\)&oldid=887219468](https://en.wikipedia.org/w/index.php?title=R_(programming_language)&oldid=887219468)