# 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**.

Fi	g. 16. C	losing the	e previoເ	us data fi	To start again						
File	Edit	Prepare	Describe	Model	Clin	File	Edit	Prepare	Describe	Model	Climatic
	New Dat	a Frame		Ct	rl+N	🗋 📂			i 🖪 📠		0
	Open Fr	om File		Ct	rl+O						
	Open Fr	om Library									
	Import f	rom ODK									
	Import f	rom CSPRC	)								
	Import f	rom Databa	ises								
	Convert						N	b Dat	ta Loa	aded	
	Save			Ct	trl+S						
	Save As										
	Export										
	Print			Ct	rl+P						
	Print Pre	view									
	Close Da	ata File				No data	loaded				

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\_datasets**.

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

#### Fig. 17 Opening the Dodoma sheet Import Dataset File: C:/Program Files (x86)// Browse New Data Frame Name: Dodoma Import Excel Options Select Sheet: Dodoma ۷ Missing Value String: Trim Trailing White Space Rows to Skip: 0 Lines to Preview: 10 🖨 Data Frame Preview: Maximum Rows To Import Year Month Day Rain Tmax ^ 1 1935 N/ Jan 1 0.0 NA 2 1935 Jan 2 6.3 NA N/ 3 1935 Jan 3 1.8 NA N/ 4 1935 Jan 4 0.0 NA N/ 5 1935 Jan 5 0.0 NA N/ 6 1935 Jan 6 0.0 NA N/ 7 1935 Jan 7 0.0 NA N/ Y < > Comment: code generated by the dialog Import Dataset Refresh Preview To Script 0k Reset Close Help

An initial objective is to provide time series graphs for the **annual mean temperatures**, both maximum and minimum . The data are daily, and have first to be averaged to an annual level. 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.

ſ	File	Edit	Prepare	Describe	Model	Climatic	Procurer	ment Tool	s	View Help	
		i 🔒 🖉	1 X 🗅 (	1. 🛛 🗍	-	0					
				D	ata Vir					Output Window	
		Year	Month	(c) Day	Raii	Tmax	Tmin	Sunh	٨	(	^
	1	1935	Jan	1	0.0	NA	NA	NA		m(Uodoma)	
	2	1935	Jan	2	6.3	NA	NA	NA		# code generated by the dialog Describe One Variable	
	3	1935	Jan	3	1.8	NA	NA	NA		col_names=c("Year", "Month", "Day", "Rain", "Tmax", "Tmin", "Sunh")), na.rm=FALSE)	
	4	1935	Jan	4	0.0	NA	NA	NA		Year Month Day Rain	
	5	1935	Jan	5	0.0	NA	NA	NA		Min. :1935 Length:28855 Min. : 1.0 Min. : 0.00	
	6	1935	Jan	6	0.0	NA	NA	NA		Median :1974 Mode :character 1st Qu.: 8.0 1st Qu.: 0.00	
	7	1935	Jan	7	0.0	NA	NA	NA		Mean :1974 Mean :15.7 Mean : 1.57	
	8	1935	Jan	8	0.5	NA	NA	NA		Max. :2013 Max. :31.0 Max. :119.80	
	9	1935	Jan	9	0.0	NA	NA	NA		NA's :91 Tmax Tmin Sunh	
	10	1935	Jan	10	0.0	NA	NA	NA		Min. :15 Min. : 8 Min. : 0	
	11	1935	Jan	11	0.0	NA	NA	NA		1st Qu.:27 1st Qu.:18 1st Qu.: 8 Median :29 Median :17 Median :10	
	12	1935	Jan	12	0.0	NA	NA	NA		Mean :29 Mean :17 Mean : 9	
		Dodoma Showii	ng 1000 c	of 28855	rows	Showing	7 of 7 c	> volumns	·	3rd Qu.:30 3rd Qu.:18 3rd Qu.:11 Max. :36 Max. :26 Max. :14 NA's :8631 NA's :8703 NA's :18451	~
ľ	Dodo	ma	-								.:

Fig. 18 The Dodoma daily data and a summary

Use the **Describe > One Variable > Summarise** dialogue.

Choose all the columns, then press OK, to produce the summaries also shown in Fig. 18.

The results include the number of missing values, and over 8 thousand values are missing for the temperature columns. (As this feature 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. The station added temperature records later.

Use the **right-click** on the **bottom tab** and choose the last option **View Data** 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.

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. We hope you enjoy, or at least tolerate, the steps below. And there is a "silver lining" at the end, as we explain in Section 4!

## 2. Preparing the data

Often the preparation stage includes calculating further columns.

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

Fig. 19. The p	With the calculate dialogue												
					(	Calcula	ations						
		Expression	Year>	1957				~					
Prepare Describe Mode	el Climatic Procurement				Logical	and Syr	mbols	¥ [	Show /	Arguments			
Data France A	8	Data Frame:			Basic					Logica	and Sy	mbols	
Data Frame	. 🔘	Dodoma	~	]	7	8	9	/	•		<	%%	[]
Check Data	W	Variables	^		4	5	6	•	^	!=	<=	%/%	•••
Column: Calculate 🕠	Calculations	Year Month		Add	1	2	3	-	Clear	1	>	:	(
Column: Generate	Duplicate Column	Day Rain			Del	0	()	+		1	>=	And	)
Column: Factor	Transform	Tmax	_	Data Options						Help			
Column: Text	Rank	Tex	1 [1] E										
Column: Date	Polynomials	IIY	ut inte	YrTemp	TALJET	ALUL IV	LUCIA						
Column: Reshape	Row Summary			Tremp				•					
		Comment:	code g	enerated by the dialo	g Calculat	tions							
		Ok	Re	set Close		Help		To Scrip	t				

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					
		Data View	Filter					
1	Year	Month (c) Dav Rain Rename Column	Data Frame:					
2	1935	Duplicate Column	N Filter:					
4	1935 1935	Convert to Factor	N Filters					
5	1935	Convert to Ordered Factor	Add Define New Filter					
7	1935 1935	Convert to Character	N N					
8	1935	Convert to Logical Convert to Numeric	N Data Options					
10	1935	Levels/Labels	Apply Options N Apply Ap					
11 12	1935 1935	Freeze to Here	N					
13	1935	Unfreeze	N Selected Ether Previous ()					
14 15	1935 1935	Sort Filter	N					
16	1935	Remove Current Filter	N Ok Reset Close Help To Script					
	1935 Jodoma	lan 1/ 0.0	N					

□ 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)

118.21	opeenying the inter						
	Define New Filter	Filter					
Data Frame: Dodoma V Variables A Day Add Tmax Tmin Surh Data Options Variable Condition	Filter By:       == v       TRUE         Add Condition       Edit Condition         Edit Condition       Remove Condition         Clear Conditions       Clear Conditions	Data Frame:       Dodoma       ✓       Filter:       Filter1         Filter1       Add       Define New Filter         Data       Options       Options         Apply Options       Apply As Filter       Apply As Subset					
New Filter Name: Filter1		Selected Hiter Preview: ((Yr1emp == 1 RUE))					
	<b>*</b>	Comment: code generated by the dialog Filter					
	Return Help	Ok Reset Close Help To Script					

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. 1st

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. 2	2. Conv	erti	ng the Year column to an c	The resulting data									
	Year	N	Month (c) Day Rain	Tmax	^		Year (o.f)	Month (c)	Day	Rain	Tmax	Tmin	^
8402	1958	43	Rename Column	28.6		8402	1958	Jan	1	0.0	28.6	18.7	
8403	1958		Duplicate Column	29.7		8403	1958	Jan	2	0.0	29.7	18.8	
8404	1958		Delete Column	29.7		8404	1958	Jan	3	0.0	29.7	17.6	
8405	1958		Convertion Factors	30.5		8405	1958	Jan	4	7.1	30.5	18.8	
8406	1958		Convert to Factor	31.2		8406	1958	Jan	5	8.9	31.2	19.2	
8407	1958		Convert to Ordered Factor	31.1		8407	1958	Jan	6	2.0	31.1	19.1	
8408	1958		Convert to Character	27.2		8408	1958	Jan	7	0.0	27.2	18.1	
8409	1958		Convert to Logical	28.9		8409	1958	Jan	8	0.0	28.9	18.8	
8410	1958		Convert to Numeric	30.0		8410	1958	Jan	9	0.0	30.0	16.7	
8411	1958		Levels/Labels	30.1		8411	1958	Jan	10	0.0	30.1	17.3	
8412	1958			31.2		8412	1958	Jan	11	0.0	31.2	19.3	
8413	1958		Freeze to Here	31.2		8413	1958	Jan	12	0.0	31.2	19.1	
8414	1958		Unfreeze	32.1		8414	1958	Jan	13	0.0	32.1	18.3	
8415	1958		Sort	31.8		8415	1958	Jan	14	0.0	31.8	18.6	
8416	1958		Filter	32.9		8416	1958	Jan	15	0.0	32.9	18.3	
8417	1958		Remove Current Filter	33.6		8417	1958	Jan	16	0.0	33.6	17.8	
8418	1958		Nemove Current Pitter	34.1	۷	8418	1958	.lan	17	0.0	34 1	19.2	$\sim$
- + F	Dodoma				>	- ( ) ( )	Dodoma						>

Fig. 22 Converting the Year column to an ordered factor

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

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



□ 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-dialo	ogue	With the resulting data								
Summaries	x	Data View								
		- 4	Year (o.f)	mean_Tmax	count_non_	mean_Tmin	^			
Summaries More Missing Options		1	1958	29.0	365	16.1				
Common		2	1959	28.7	365	16.3				
N Non Missing N Total		3	1960	29.0	365	15.9				
N Missing Mode		4	1961	29.3	365	17.1				
All but (upordered) Factor		5	1962	29.0	365	16.1				
	m	6	1963	28.5	363	16.0				
		7	1964	28.9	360	15.7				
		8	1965	28.8	363	16.0				
Numeric		9	1966	29.1	365	16.6				
Sum Median		10	1967	28.5	365	16.7				
Mean Standa	rd Deviation	11	1968	27.9	366	15.6				
Variance		12	1969	29.7	365	17.0				
Quartiles		13	1970	28.6	365	16.5				
	Quartile	14	1971	28.5	365	16.3				
		15	1972	28.8	366	16.6				
		16	1973	29.5	362	16.6				
		17	1974	28.8	304	16.2	~			
Return Help		SI	howing 56	ona_oy_rear	s   Showir	ng 5 of 5				

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
Duplicate Column	Data View
Data Frame: Dodoma_by_Year ▼ Variables Year mean_Tmax count_non_missing_Trr mean_Tmin Mad Column to Duplicate: Year Position of Duplicated Column ○ Beginning ○ Before	Year (o.f)         Year1         mean Tmax         count non         A           1         1958         1958         Rename Column         Duplicate Column           2         1959         1959         Duplicate Column         Duplicate Column           3         1960         1960         Convert to Factor
Count_nor_missing_in     Data        • After          • Options         • End           • New Column Name: Year1         • Comment: code generated by the dialog Duplicate Column	5         1962         1962           6         1963         1963           7         1964         1964
Ok Reset Close Help To Script	8 1965 1965 Convert to Logical Convert to 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	g. 26. The li	ne plot menu	And the dialogue					
Descr	ibe Model	Climatic	Procurement Tools View	Line Plot					
1	One Variable Two Variables Three Variable	+ + !5 +	ata View	Data Frame: Dodoma_by_Year V Variables Mean_Tmax					
5	Specific	•	Frequency Tables	Year					
(	General	•	Summary Tables	Year1 nean_Tmax					
1	Multivariate	•	Multiple Response	count_non_missing, mean_Tmin V Data					
ī	Jse Graph		Scatter Plot	< > Options X Variable:					
(	Combine Grap	ohs	Line Plot	Line Options Year1					
1	Themes		Histogram	Options Factor (Optional):					
۱ ا	/iew Graph		Boxplot	l ✓ Points					
 م	20.0	2005	Dot Plot	✓ Add Line of Best Fit ✓ With Standard Error					
0	29.1	365	Rug Plot	Save Graph					
/	28.5	365	Bar Chart	Comment: code generated by the dialog Line Plot					
8 9	27.9 29.7	365 365	Cumulative Distribution	Ok Reset Close Help To Script					

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



## 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.

File Edit Prepare Describe Mc New Data Frame Ctrl+N Open From File Ctrl+O	odel Climatic Procurement	Save Data As
Open From Library	Data View	
Import from ODK	max count_non_ mean_Tmin 365 16.1 3	Save Data To: C:/Users/Roger/Documents/R-Instat/Dodoma Tutorial RDS Browse
Import from CSPRO	365 16.3 3	
Import from Databases	365 15.9 3	Click Ok to confirm the save
Convert	365 16.1 3	
Save Ctrl+S	363 16.0 3	Community and a second of builton dialog Court Date As
Save As 🕨 🕨	Save Data As	Comment: code generated by the dialog Save Data As
Export •	Save Output Window As	
Print Ctrl+P	Save Log As	Ok Reset Close Help To Script
Print Preview	Save Script Window As	
Close Data File	366 15.6 3	

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 <u>issues page</u> on our <u>GitHub</u> 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: <u>https://github.com/africanmathsinitiative/R-Instat/issues</u>. 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