BE540 - Introduction to Biostatistics
Computer Illustration

Topic 1 – Summarizing Data
Software: SPSS

A Visit to Yellowstone National Park, USA

Source:
Chatterjee, S; Handcock MS and Simonoff JS *A Casebook for a First Course in Statistics and Data Analysis.* New York, John Wiley, 1995.

Setting:
Upon completion of BE540, you decide to take a vacation to the United States. Of particular interest is seeing an eruption of the famous "Old Faithful" geyser at Yellowstone National Park. Unfortunately, your time is limited and you do not wish to miss seeing an eruption.

This worked example illustrates descriptive analysis of a data set of 222 interval times between eruptions of the Old Faithful Geyser, measured during August 1978 and 1979.

Data File:
GEYSER1.DAT - This is a data set in ASCII format.

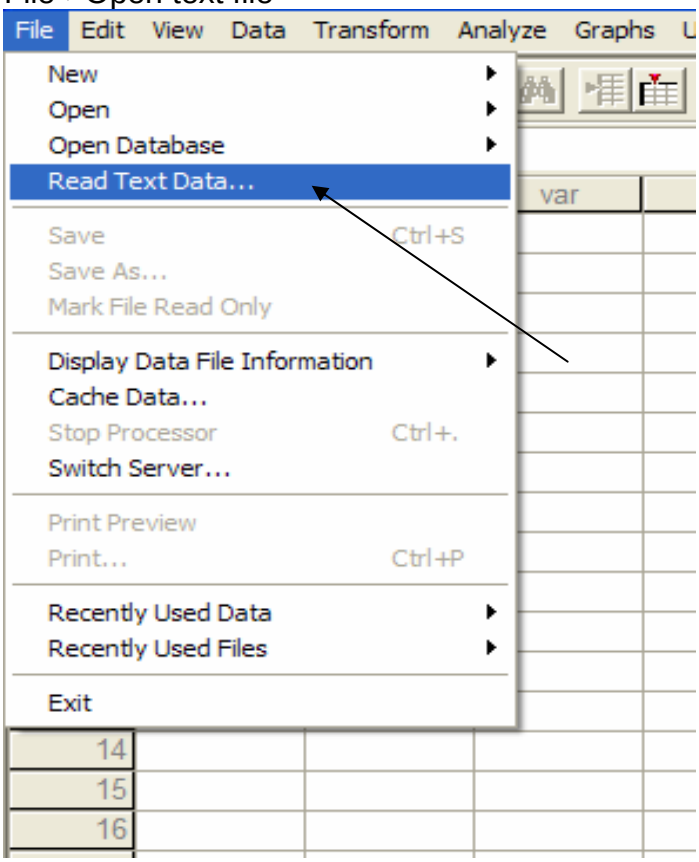
Description of Data:
There are three variables, in the following order:

> INDEX - An index of the date of the eruption. We will not be using this variable.
>
> DURATION - The duration of the eruption in minutes.
>
> INTERVAL - The length of the interval between the current eruption and the next eruption.

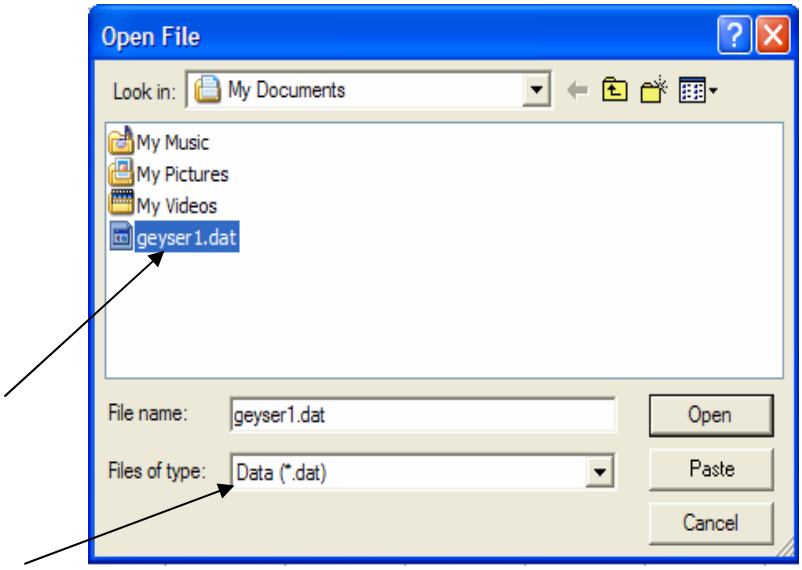
Objective:
Describe the pattern of eruptions and predict the interval of time to the next eruption.

1. Read in the ASCII format data 'GEYSER1.DAT':

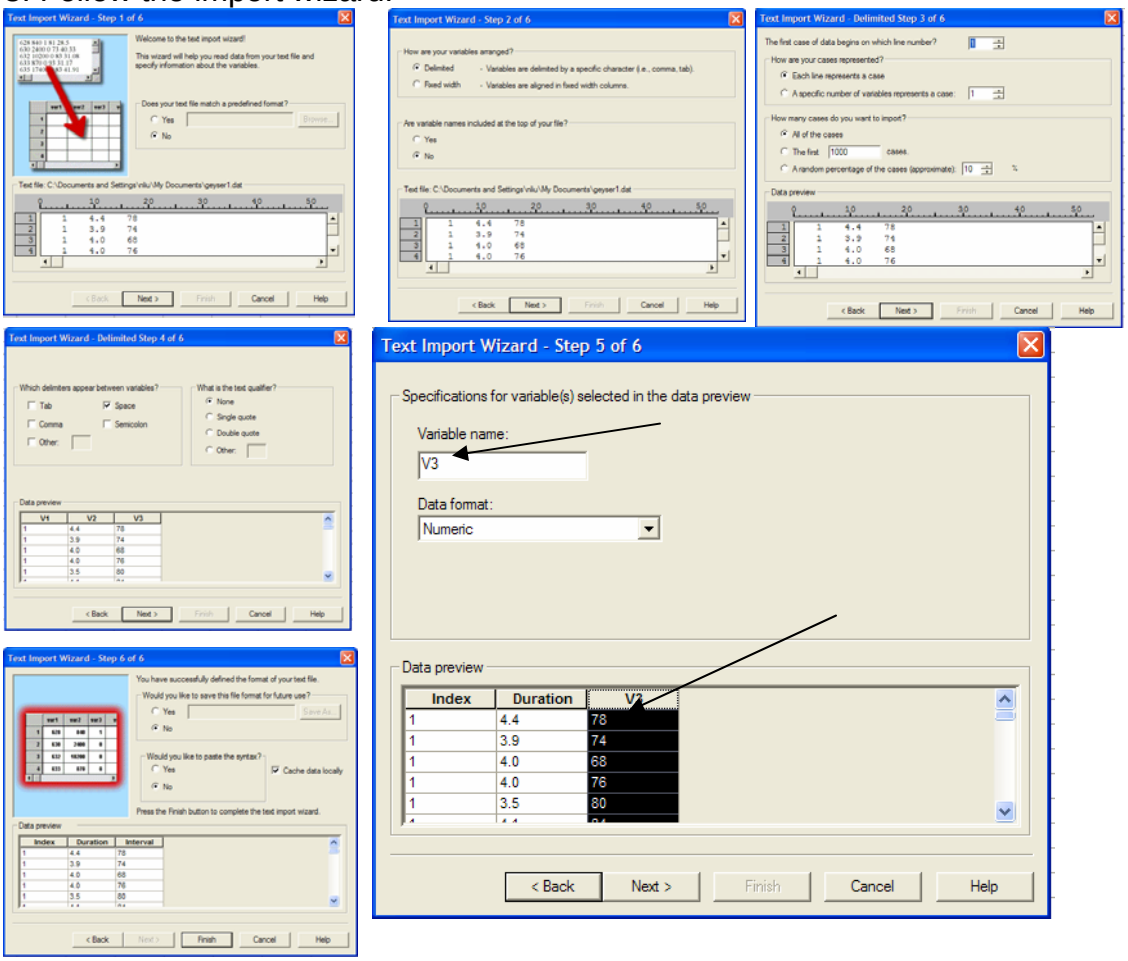
1.
File->Open text file



2. Choose the data file:



3. Follow the import wizard:

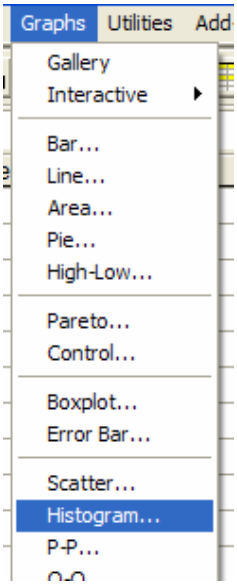


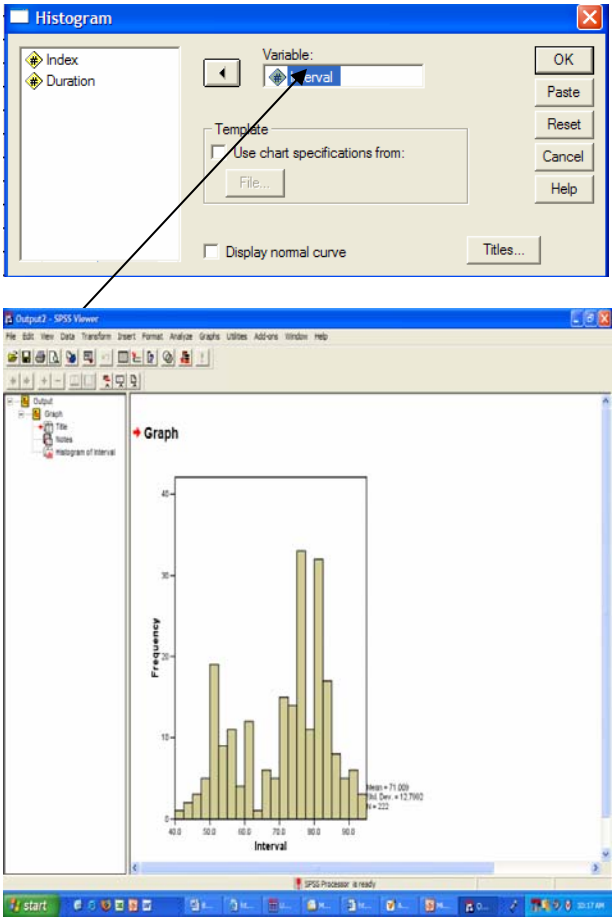
Data set (222 observations) is imported to SPSS. You should see the following worksheet:

	Index	Duration	Interval
1	1	4.40	78.0
2	1	3.90	74.0
3	1	4.00	68.0
4	1	4.00	76.0

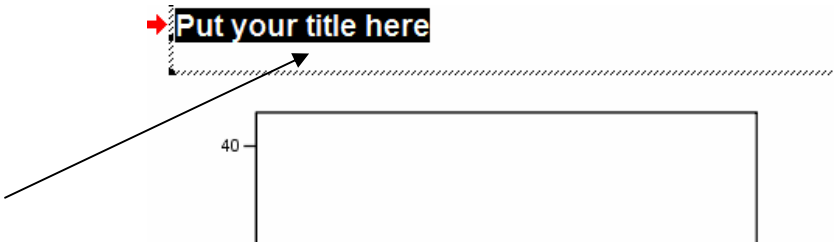
2. Obtain a Histogram of Interval Times.

Graphs>Histogram…>Select variable>Click OK

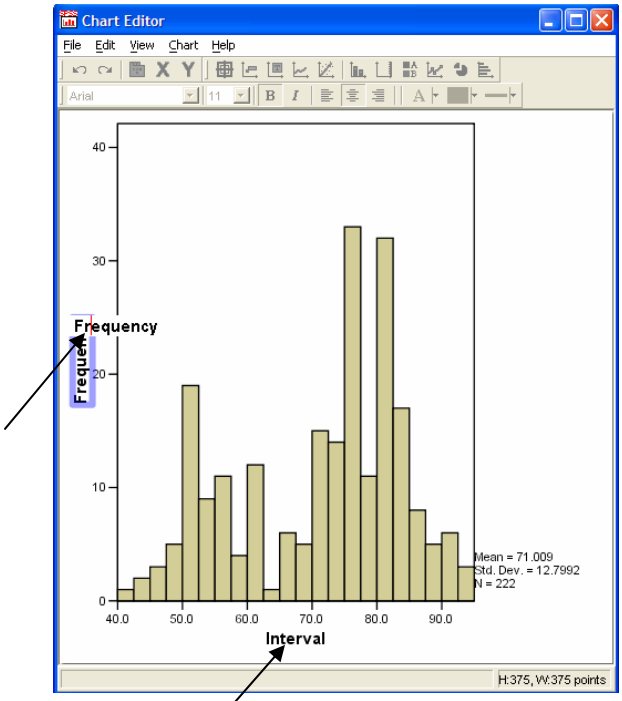




This histogram is generated with SPSS default settings. If you like, you can revise the format by simply double clicking on the part you want to change. The following are some examples of how to change graph options.

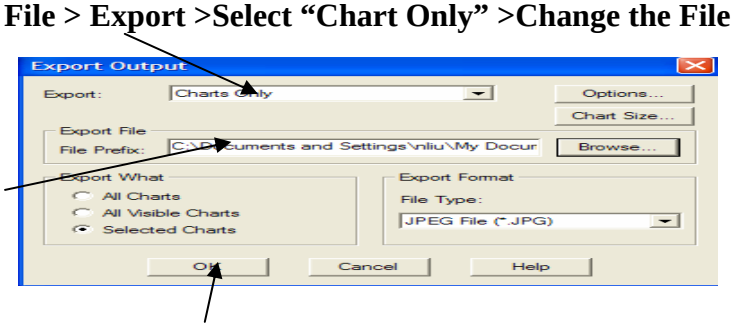


Double-Click the chart to active "Chart Editor". You can change X-label and Y-label by click them.



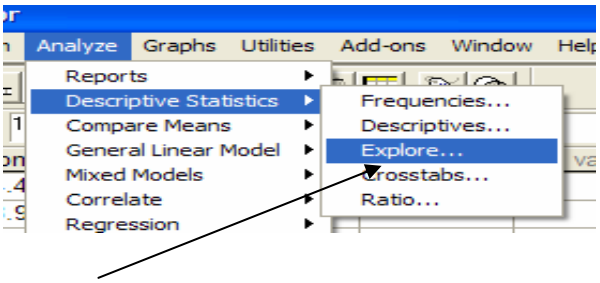
3. Save this histogram as a picture that you can print directly or that you can insert into a document such as this one.

File > Export >Select "Chart Only" >Change the File name if needed>Click OK



4. Instead of a histogram, we might have constructed a stem-and-leaf diagram.

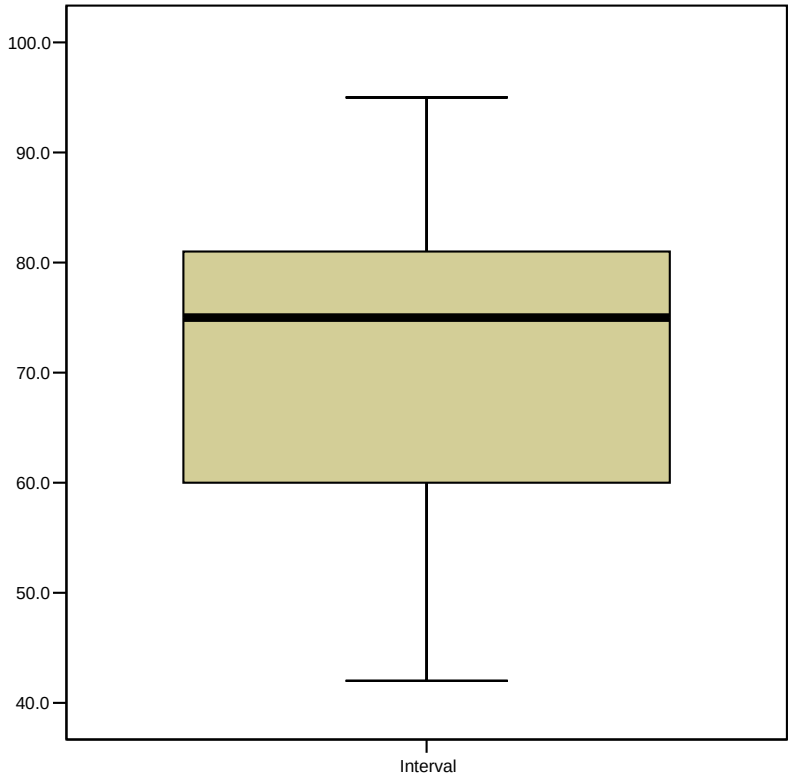
Analyze>Descriptive Statistics>Explore…>Choose the variable>Click “OK”.




```
Interval Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00        4 .  &
     8.00        4 .  589&
    28.00        5 .  0111111123344
    15.00        5 .  556778&
    13.00        6 .  00112&
    11.00        6 .  6789
    29.00        7 .  001122333344
    44.00        7 .  555555566666777788899
    49.00        8 .  00000011112222233334444
    13.00        8 .  66689&
     8.00        9 .  01&
     1.00        9 .  &

 Stem width:  10
 Each leaf:       2 case(s)


 & denotes fractional leaves.
```

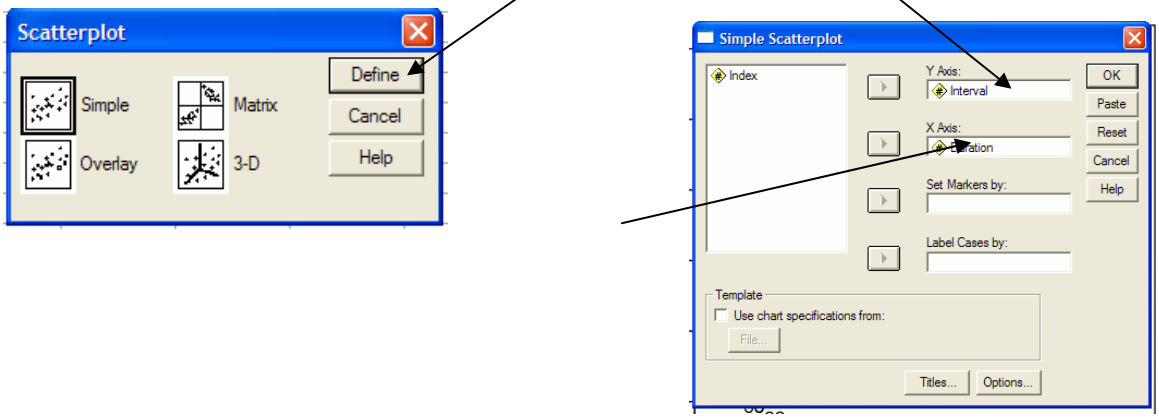
5. In this example, a Box and Whisker plot is not very informative. Let's see why.
The same SPSS instruction (Analyze>Descriptive Statistics>Explore…>Choose the variable>Click "OK".) illustrated above also yields a Box and Whisker plot.

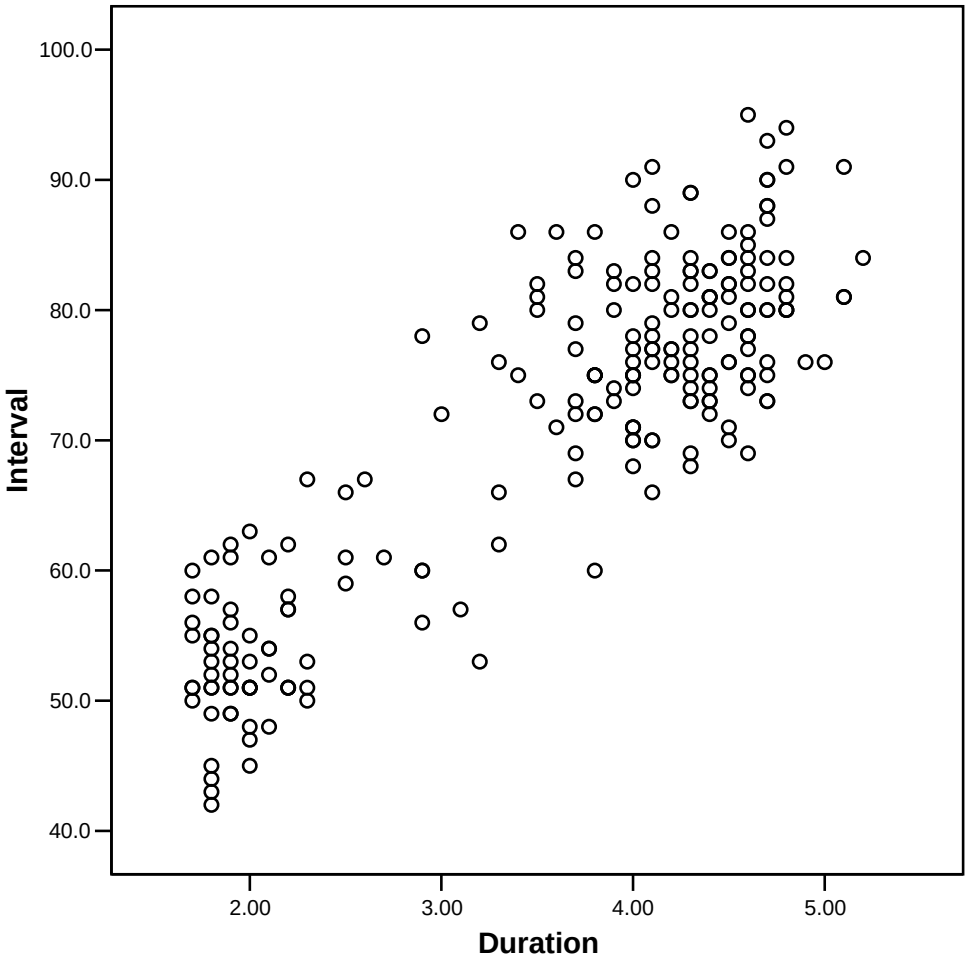
```
Interval Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00        4 .  &
     8.00        4 .  589&
    28.00        5 .  0111111123344
    15.00        5 .  556778&
    13.00        6 .  00112&
    11.00        6 .  6789
    29.00        7 .  001122333344
    44.00        7 .  555555566666777788899
    49.00        8 .  00000011112222233334444
    13.00        8 .  66689&
     8.00        9 .  01&
     1.00        9 .  &

 Stem width:  10
 Each leaf:       2 case(s)


 & denotes fractional leaves.
```



6. We have information on duration of eruption also. One possibility is that the duration of the current eruption is a predictor of the interval time to the next eruption. To investigate this possibility, construct a scatter plot of interval time versus duration. Plot the predictor DUR on the horizontal axis (X) and the outcome INTERVAL time to the next eruption on the vertical axis (Y).

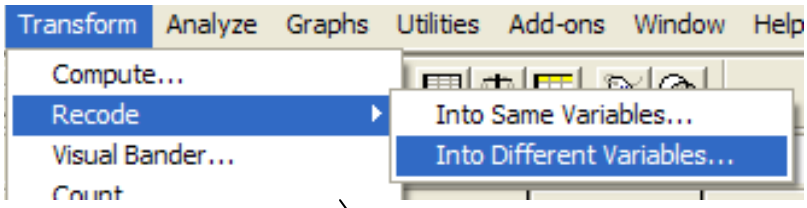
Graph>Scatter>Define>Select "X axis" and "Y axis" >click "OK"

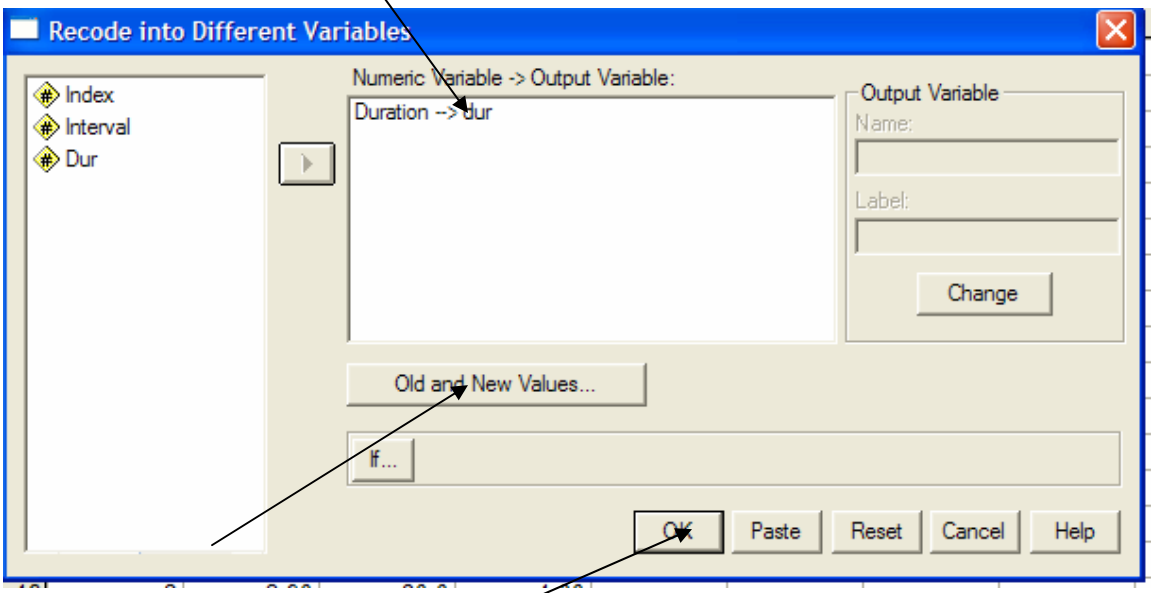




7. Create a grouped measure of duration and construct separate box and whisker plots of interval times for the interval times that follow eruptions less than 3 minutes in duration and the interval times that follow eruptions greater than 3 minutes in duration.

First, we need to recode the variable.
Transform>Recode>Recode to a different variable>Select the variable define the recode rule>Click "OK".

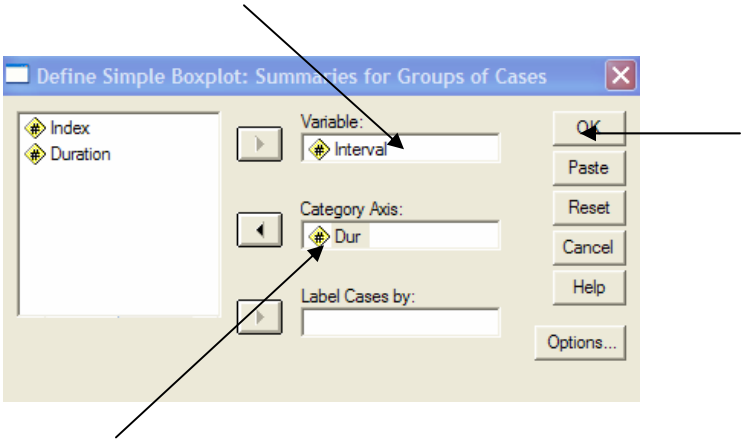




1 : Dur

	Index	Duration	Interval	Dur
1	1	4.40	78.0	1.00
2	1	3.90	74.0	1.00
3	1	4.00	68.0	1.00
4	1	4.00	76.0	1.00
5	1	3.50	80.0	1.00
6	1	4.10	84.0	1.00
7	1	2.30	50.0	.00

Then Graphs>Boxplot…>Simple>define>Select Variable and Category >Click "OK"



You should see.