MIS771 Descriptive Analytics and Visualisation	DEAKIN UNIVERSITY
Topic 8 Tutorial - Basics of Data Visualisation	

Table of Contents

Introduction

This week's tutorial provides an introduction to the Tableau Data Visualisation tool, explains how to open/connect to a data set, and what elements are required to generate a basic Data Visualisation. You will also practice adding simple trend lines to charts.

Specifically, the aims of this tutorial are to:

• Understand the basic elements of the Tableau Data Visualisation tool
• Understand how basic data structures are used as Inputs for Data Visualisation (Categorical, Continuous, and Ordinal)
• Represent a simple set of data points in a series of Scatter Plot charts

1. Scenario

We will begin with the analysis of the Anscombe's quartet data set.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

The main aim of today's tutorial is to see how using a simple chart (in this case a Scatter Plot) can reveal patterns in the Anscomebe's data that are not obvious by looking at the tabular values alone.

2. Open the data file in Excel

First we need to download our sample data set for this Tutorial and open the data in Excel.

a) Download the file Anscombe_Tut04.xls from the Unit Resources on Cloud Deakin. Save it to your hard drive (or your working disk if working on one of the Deakin lab computers).

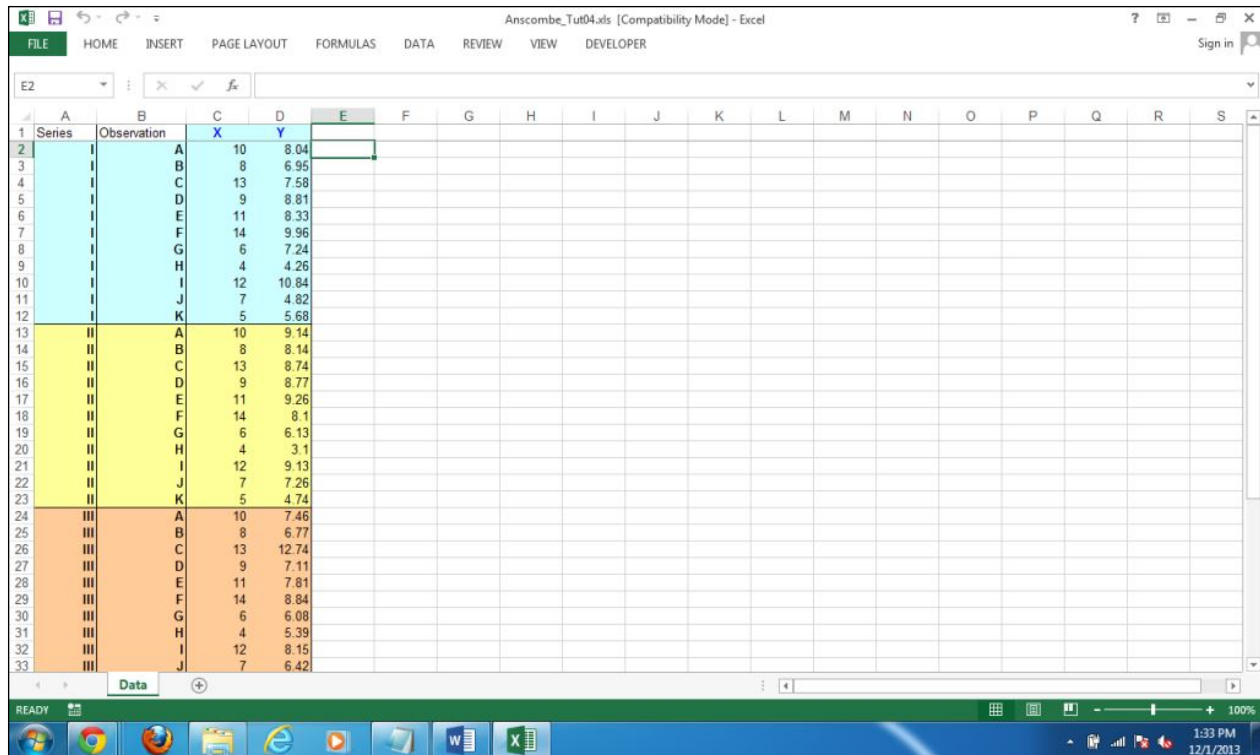b) Open the Anscombe_Tut04.xls file in Excel.

Figure 1: The Anscombe_Tut04.xls file, opened in Excel

3. What structure can you see in that data?

Before we can begin any type of analysis on the Anscombe data we first need to understand the basic structure of the data and understand how the data is organised.

In the lecture you have been shown the difference between Categorical and Continuous data. In this step we will first identify the different types of Categorical and Continuous data available in the tutorial data set.

> ⚠️ Categorical Data usually can take on one of a limited, and usually fixed, number of possible values, and is typically used to describe or group related data. Continuous data is usually Numeric and many take on any value with a range.

a) What different groups of data are there in the Data Set, and what are the labels/names for each of the Categorical data?

b) What are the labels/names for the Continuous data?

c) How many separate data series are there?

d) How many observations are there in each series?

4. What relationships can you see in that data?

In this exercise we will examine the Anscombe data and performing some simple Statistical Analysis across each of the Data Series (such as Mean, Variance, and Correlation) in a tabular format using Excel.

- What is the Mean of the y data for each Series, and how do they compare?
- What is the Variance of the x data for each Series, and how do they compare to each other?
- What is the Correlation between x and y data in each Series?

It is assumed you know how to create simple formulas in Excel. In this step we will add some basic Excel formulas to the tutorial spreadsheet and review some basic analysis across each data series.

Figure 2: Result after performing some basic data analysis on the Anscombe data

(a) Compare the Mean of the x values across each Series

- In Cell G3 enter the formula to calculate the Mean on all the x values in Series I
- In Cell G4 enter the formula to calculate the Mean on all the x values in Series II
- In Cell G5 enter the formula to calculate the Mean on all the x values in Series III
- In Cell G6 enter the formula to calculate the Mean on all the x values in Series IV

> ⚠ Excel uses the formula 'AVERAGE' for calculating the Mean of a set of data. For example, to calculate the Mean or Average of all the x data in Series I the excel formula would be '=AVERAGE(C2:C12)'.

(b) Compare the Mean of the y values for each Series.

- In Cell H3 enter the formula to calculate the Mean on all the y values in Series I
- In Cell H4 enter the formula to calculate the Mean on all the y values in Series II
- In Cell H5 enter the formula to calculate the Mean on all the y values in Series III
- In Cell H6 enter the formula to calculate the Mean on all the y values in Series IV

⚠ Excel uses the formula 'AVERAGE' for calculating Mean of a set of data. For example, to calculate the Mean or Average of all the y data in Series III the excel formula would be '=AVERAGE(D24:D34)'.

(c)   Compare the Variance of the x values across each Series.

- In Cell I3 enter the formula to calculate the Variance on all the x values in Series I
- In Cell I4 enter the formula to calculate the Variance on all the x values in Series II
- In Cell I5 enter the formula to calculate the Variance on all the x values in Series III
- In Cell I6 enter the formula to calculate the Variance on all the x values in Series IV

⚠ Excel uses the formula 'VAR' for calculating the Variance across a set of data. For example, to calculate the Variance across all the x data in Series IV the excel formula would be '=VAR(C35:C45)'.

(d)   Compare the Variance of the y values across each Series.

- In Cell J3 enter the formula to calculate the Variance on all the y values in Series I
- In Cell J4 enter the formula to calculate the Variance on all the y values in Series II
- In Cell J5 enter the formula to calculate the Variance on all the y values in Series III
- In Cell J6 enter the formula to calculate the Variance on all the y values in Series IV

⚠ Excel uses the formula 'VAR' for calculating the Variance across a set of data. For example, to calculate the Variance across all the y data in Series II the excel formula would be '=VAR(D13:D23)'.

(e)   Compare the Correlation between x and y data in each Series.

- In Cell K3 enter the formula to calculate the Correlation between all x and y values in Series I
- In Cell K4 enter the formula to calculate the Correlation between all x and y values Series II
- In Cell K5 enter the formula to calculate the Correlation between all x and y values Series III
- In Cell K6 enter the formula to calculate the Correlation between all x and y values Series IV

⚠ Excel uses the formula 'CORREL' for calculating Correlation between two set of data. For example to calculate the Correlation between all the x and y values in Series II the excel formula would be '=CORREL(C13:C23,D13:D23)'.

**5. Starting Tableau and connecting to our sample data.**

In this exercise we will generate a Data Visualisation of the Anscombe data by connecting to the data using the Tableau Software and representing the data in a series of charts (using Scatter Plot charts).

- How do we start the Tableau software?
- How do we connect to the Data?
- How do we generate a Chart?
- How do we add a Trending line?

- How can we compare different data Series?

In the remainder of this tutorial we will work through the steps of starting the Tableau software, connecting to the Anscombe data set, and configuring the Tableau workspace to display the data as a series of Scatter Plot charts.

First start the Tableau software, just like any other software (access from the Windows Start menu)… Note: If you haven't already you can download a trail version of Tableau from http://www.tableau.com/ and use the Product Activation Key provided on Cloud Deakin.
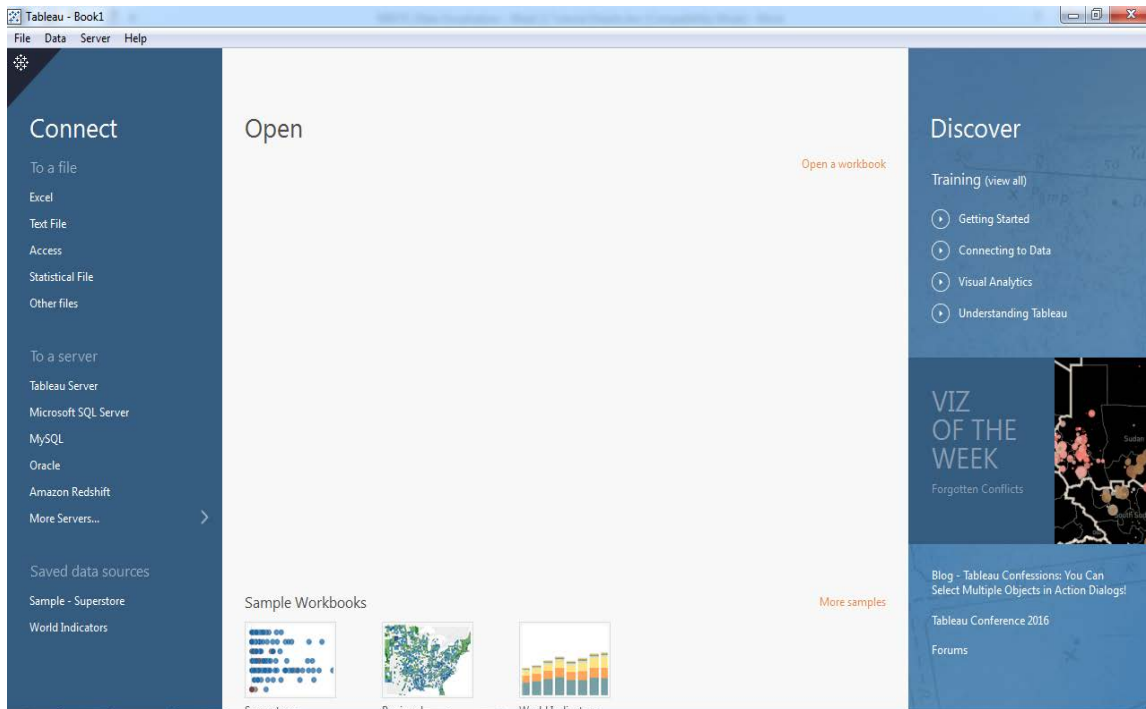
Figure 3: The Tableau software opening screen (version 9.3 shown)

> ⚠ If you are using the Deakin Lab computers the Tableau software will already be installed. If you are working through the tutorial on your personal computer you will need to first install the Tableau using the instructions and software key provided in the MIS771 discussion forum.

a) Next use the 'Connect' feature in the left panel to open the Anscombe_Tut04.xls in Tableau.
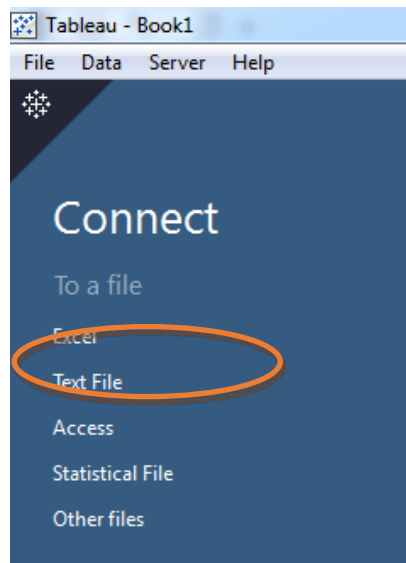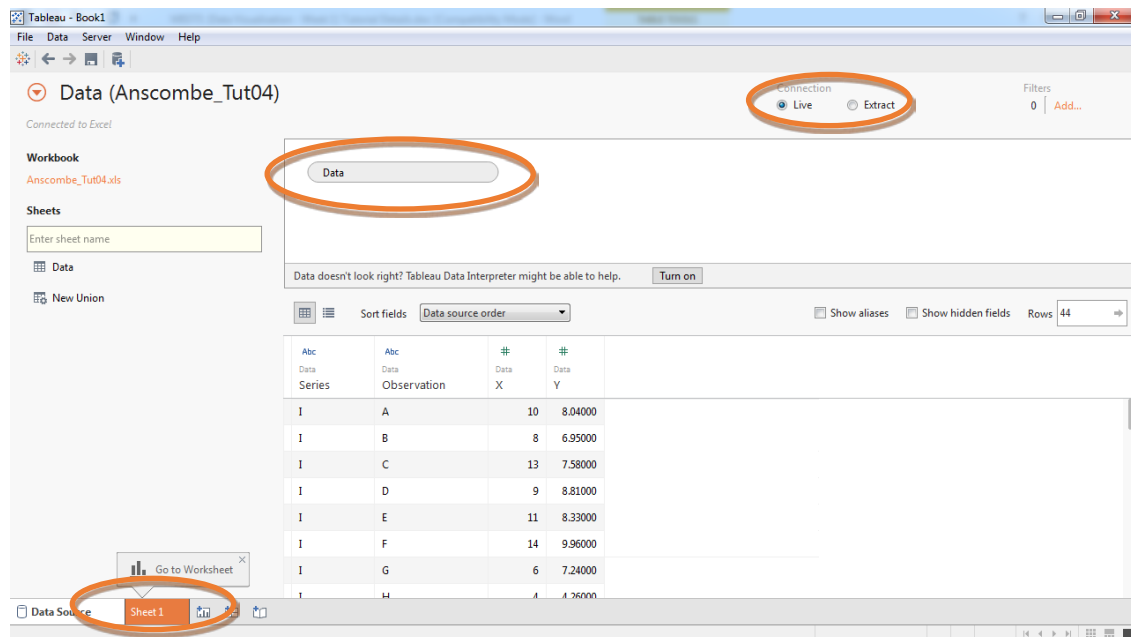
| Step 1. 'Select Excel'	Step 2. Locate and Select 'Anscombe_Tut04.xls' and click 'Open'
Figure 4: Tableau Connect to Data	*Figure 6: Select the Anscombe Tutorial file*
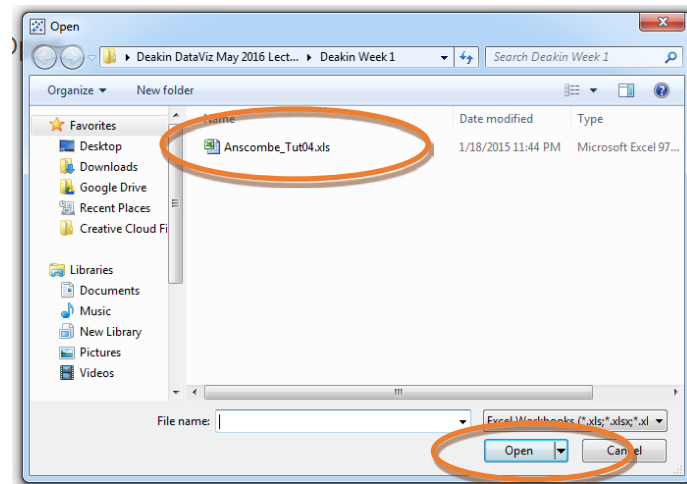
Step 3. 'Connect live'. Ensure the 'Live' Connection option is selected, and that the 'Data' worksheet is selected and displaying in the Preview, then click the 'Sheet 1' tab at the bottom of the screen.

Figure 8: Tableau Connect live to the Excel file

b) Can you identify the labels for the Categorical data and the Continuous data?
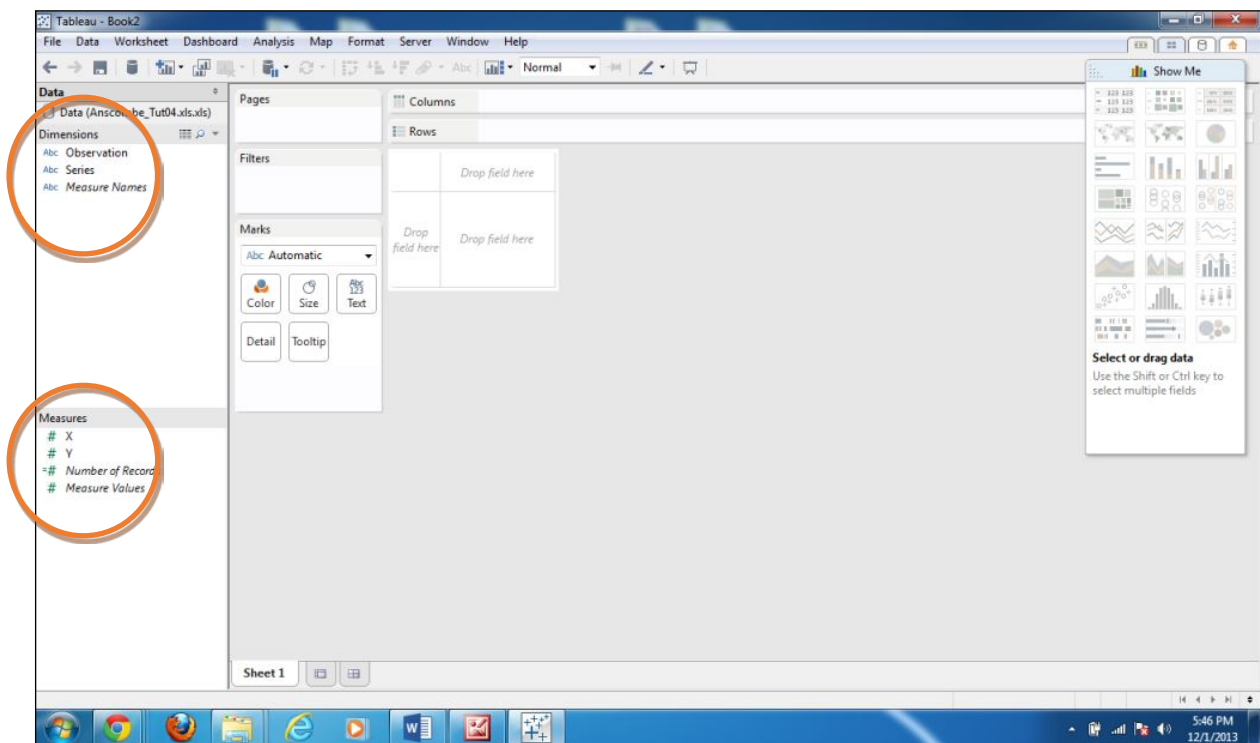
Figure 9: Tableau Work Sheet

Tableau calls Categorical data 'Dimensions', and calls Continuous data 'Measures'. Note: if you saved the Anscombe_Tut04.xls with your additional Data Analysis from Step 5. there may be some additional unnecessary Dimensions (for example, Average), these can be hidden in Tableau, using the 'Hide' option on each unwanted Dimension
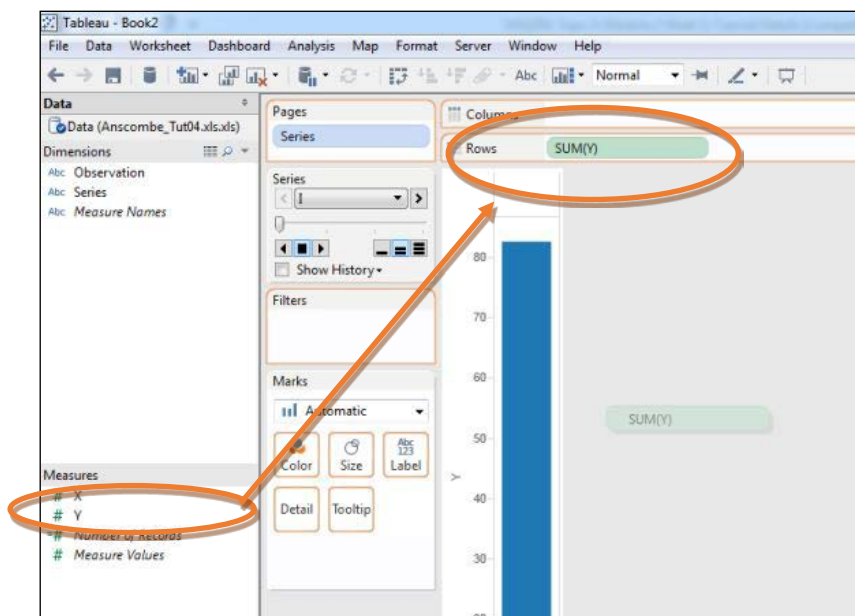
6. **Configuring a Tableau Workspace to build a Chart.**

    a) Begin by placing the 'Series' Category/Dimension on to the Pages Panel (drag-and-drop).
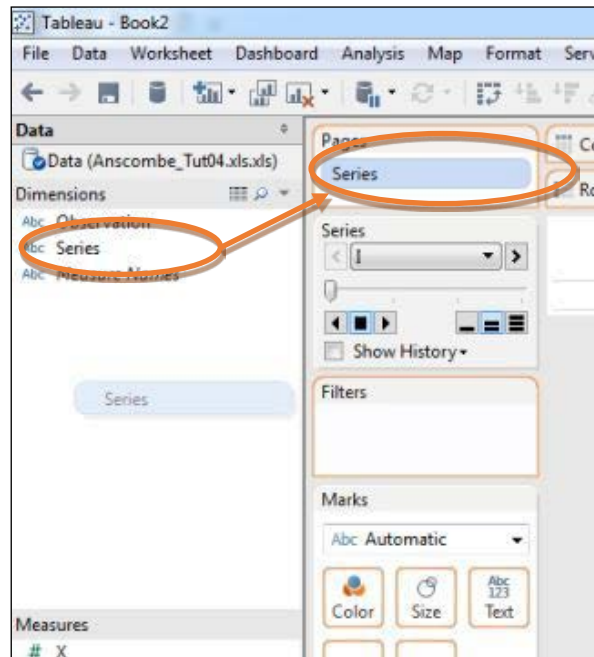
Figure 12: Drag-and-drop 'Series' to the Pages Panel

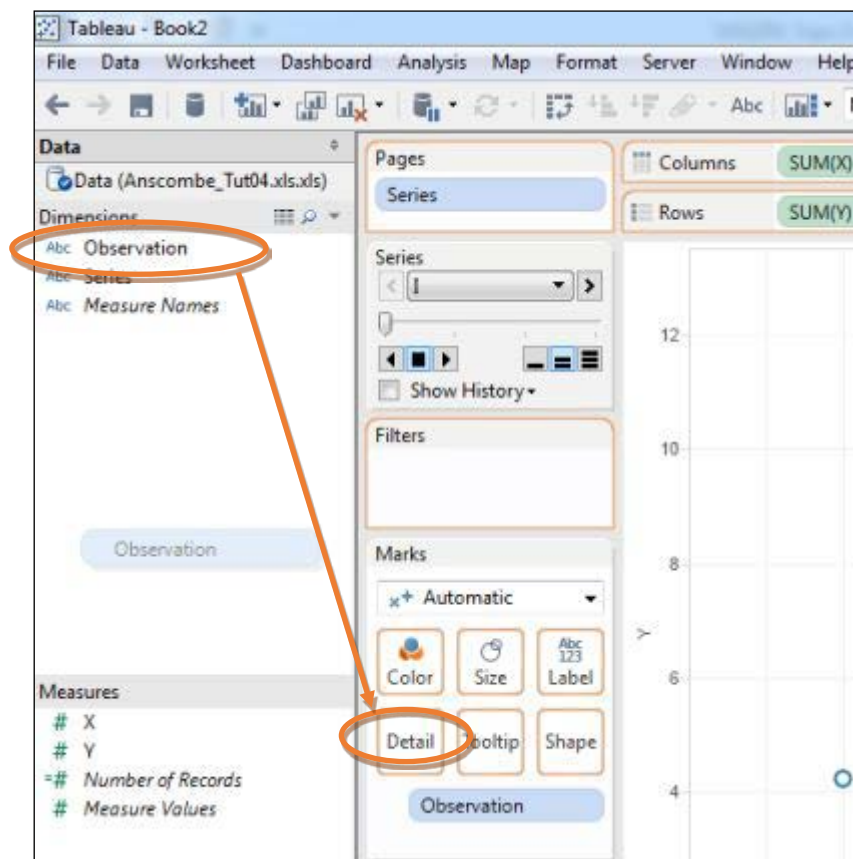   b) Place the 'Y' Measure on to the Rows Panel (drag-and-drop).

Figure 13: Drag-and-drop 'Y' to the Rows Panel

c) Place the 'X' Measure on to the Columns Panel (drag-and-drop).
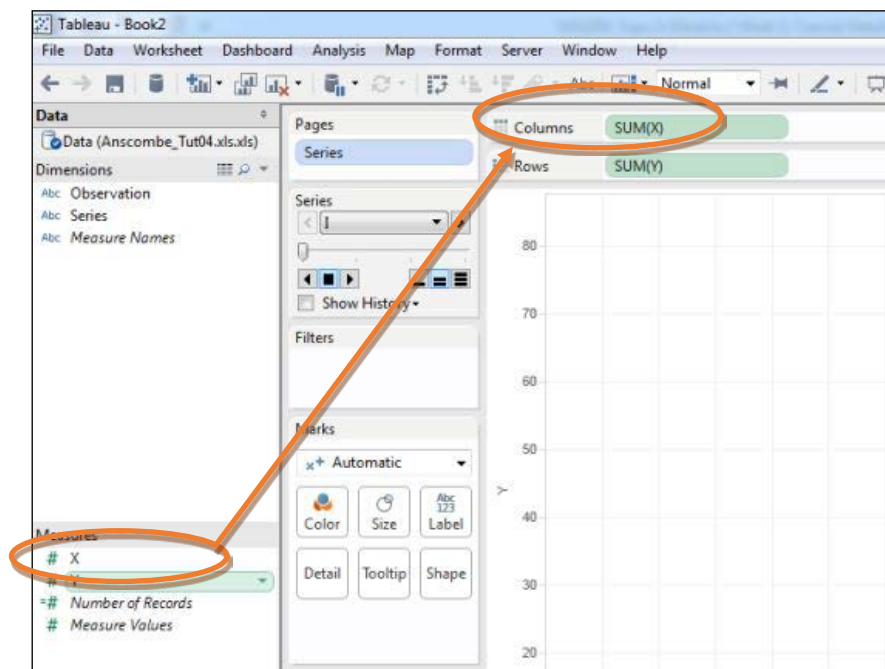
Figure 14: Drag-and-drop 'X' to the Columns Panel

d) Finally place the 'Observation' Category/Dimension on to the Detail Panel (drag-and-drop).

Figure 15: Drag-and-drop 'Observation' to the Detail Panel

## 7. Adding additional elements to the Data Visualisation to revel additional insights

We can also provide additional visual elements to our chart by applying a built-in Liner Regression algorithm to the Data we have in our Anscombe_Tut04.xls data file.

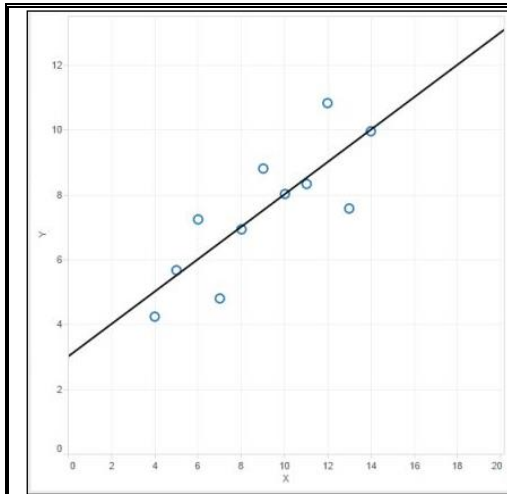a) Add a Trend Line to the chart area (right-click on the chart).
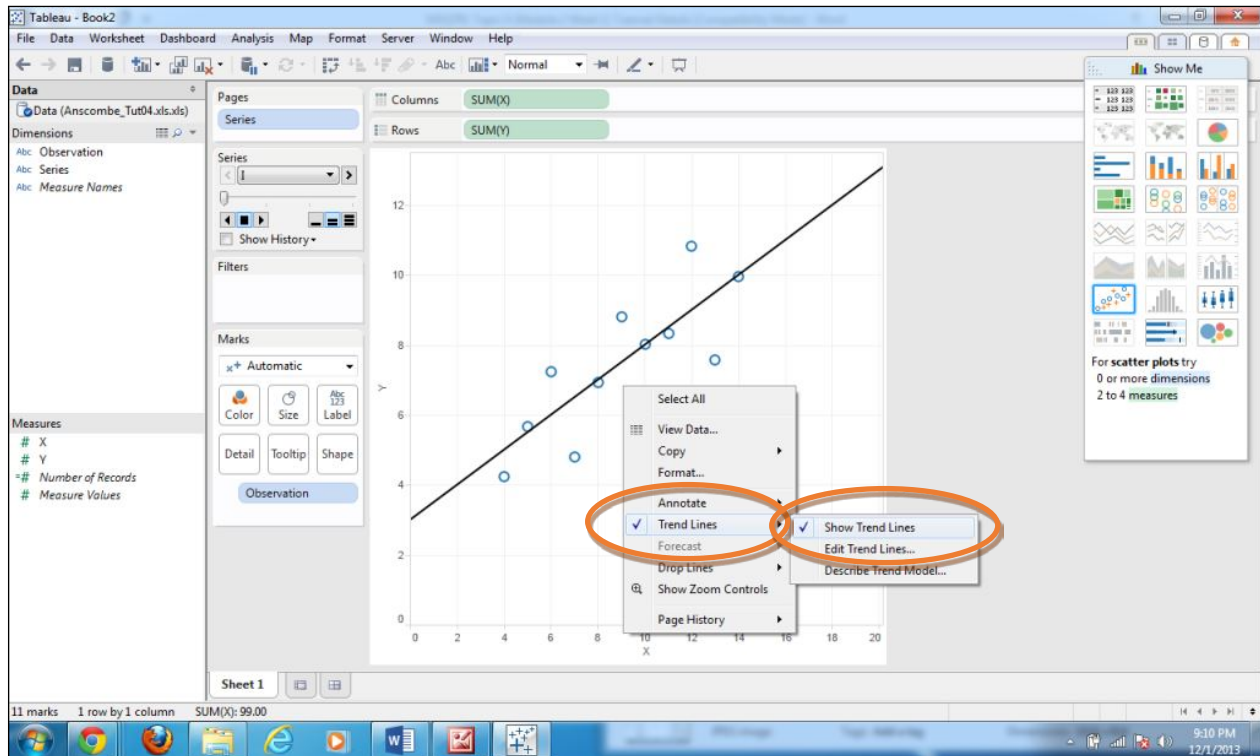
Figure 16: Drag-and-drop 'X' to the Columns Panel

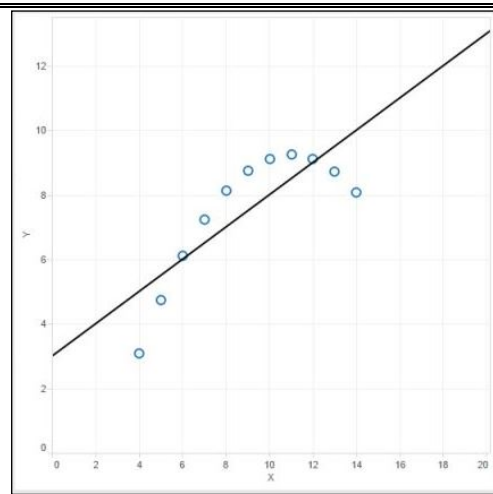b) Use the Next and Previous buttons on the Pages Panel to review each Series.
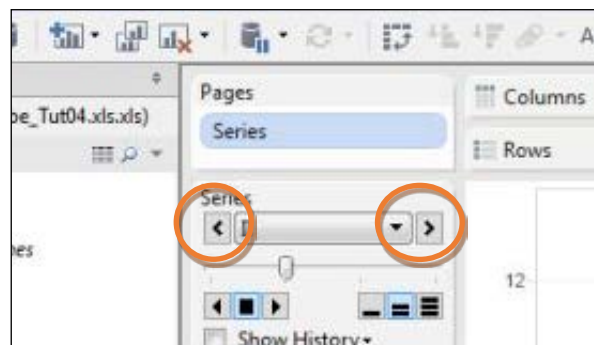
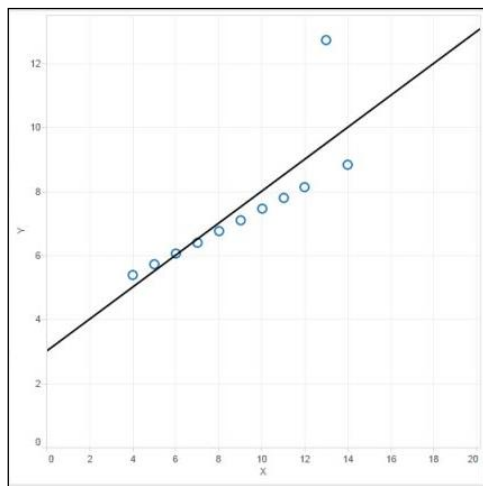*Figure 17: Use the Next and Previous buttons to review each Series*

c) Use the Next and Previous buttons on the Pages Panel to review each Series one at a time on the screen.

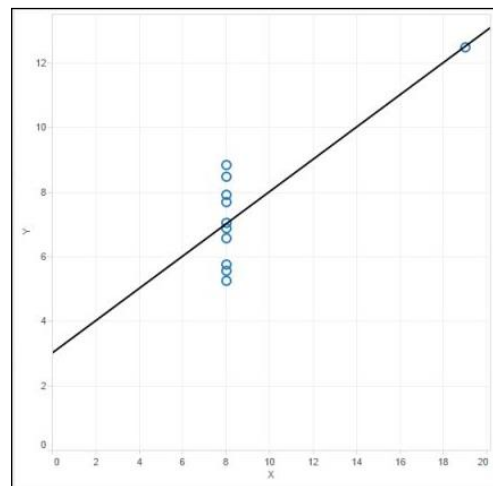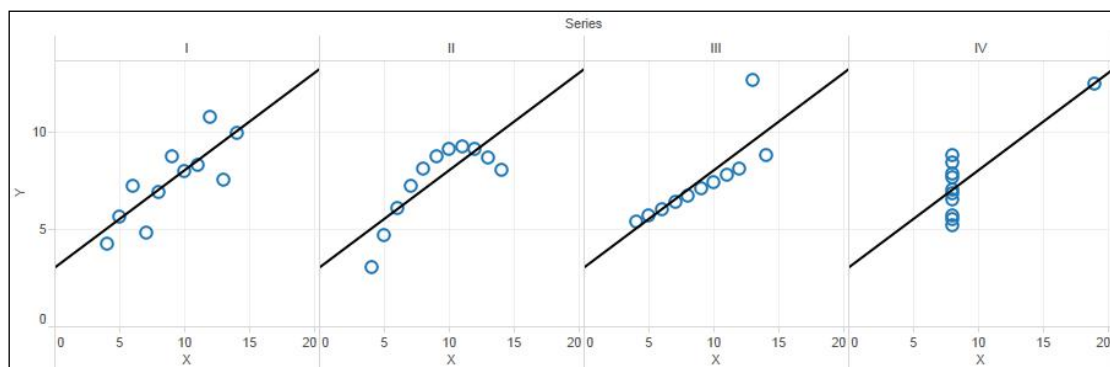*Figure 18: Series I*

Figure 19: Series II

Figure 20: Series III

Figure 21: Series IV

d) Next, can you work out how to display all four charts at the same time?

Figure 22: Displaying all four series on the same sheet.

Try moving (by dragging-and-dropping) the `Series` dimension from the 'Pages' shelf to the 'Columns' shelf. Also try moving it to the 'Rows' shelf and see if it has any effect of the display layout.

8. **Alternative Charts and Exporting to PowerPoint**

   a) Next by yourself, add a new Worksheet to your Tableau Workbook and experiment with different chart types to display the Anscombe_Tut04.xls data (for example, try using Bar Charts or Box-and-whisker Charts)?

On the new Worksheet 'drop-and-drag' various combinations of Dimensions and Measures to the 'Columns', 'Rows', and 'Marks' shelfs, then use the 'Show Me' feature to select different chart types.

Note: After selecting a new Chart type you may need to experiment with the placement of the various Dimensions and Measures to get the chart displaying the way you'd like it to.

   b) Finally, using the 'Export -> Image' option on the Tableau 'Worksheet' menu, save one or more of your worksheet displays out of Tableau and import them in to a Microsoft PowerPoint presentation.

9. **Effectiveness of Data Visualisations**

Using your insights from this Tutorial, briefly summarise what you've learnt about the Anscombe data set? Do you think using a Data Visualisation helped you understand the data more effectively?

References

1. Source Material: Francis John "Frank" Anscombe (13 May 1918 – 17 October 2001) was an English statistician
2. Source of data: http://web.uvic.ca/~dgiles/blog/anscombe.xls (with some manipulation)
3. Tool: Tableau Version 8.1 http://www.tableausoftware.com/